Entomology in Focus 3

Chandrasekar Raman
Marian R. Goldsmith
Tolulope A. Agunbiade *Editors*

Short Views on Insect Genomics and Proteomics

Insect Genomics, Vol.1

SOCIEDADE ENTOMOLÓGICA DO BRASIL · 1972 ·

Springer

Entomology in Focus

Volume 3

Series editor
Fernando L. Cônsoli, Piracicaba, São Paulo, São Paulo, Brazil

This series features volumes of selected contributions from workshops and conferences in all areas of current research activity in applied mathematics. Besides an overall evaluation, at the hands of the publisher, of the interest, scientific quality, and timeliness of each proposal, every individual contribution is refereed to standards comparable to those of leading mathematics journals. This series thus proposes to the research community well-edited and authoritative reports on newest developments in the most interesting and promising areas of mathematical research today.

More information about this series at http://www.springer.com/series/10465

Chandrasekar Raman • Marian R. Goldsmith
Tolulope A. Agunbiade
Editors

Short Views on Insect Genomics and Proteomics

Insect Genomics, Vol.1

Springer

Editors
Chandrasekar Raman
Department of Biochemistry
and Molecular Biophysics
Kansas State University
Manhattan, KS, USA

Marian R. Goldsmith
Biological Sciences Department
University of Rhode Island
Kingston, RI, USA

Tolulope A. Agunbiade
Department of Internal Medicine
Yale University School of Medicine
New Haven, CT, USA

Howard Hughes Medical Institute
Chevy Chase, MD, USA

ISSN 2405-853X ISSN 2405-8548 (electronic)
Entomology in Focus
ISBN 978-3-319-24233-0 ISBN 978-3-319-24235-4 (eBook)
DOI 10.1007/978-3-319-24235-4

Library of Congress Control Number: 2015957423

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.com)

Short Views on Insect Genomics and Proteomics, Volumes 1 and 2

Editor information

Dr. Chandrasekar Raman
Department of Biochemistry and Molecular Biophysics
Kansas State University
Manhattan KS 66506
USA
Tel. 1-785-532-6125; mobile: 1-859-608-7694
Email: biochandrus@yahoo.com; chandbr@ksu.edu

Prof. Marian R. Goldsmith
Biological Sciences Department
University of Rhode Island
CBLS – 120 Flagg Road
Kingston RI 02881
USA
Tel. 1-401-874-2637
Email: mrgoldsmith@uri.edu

Dr. Tolulope A. Agunbiade
Section of Infectious Diseases
Department of Internal Medicine
Yale University School of Medicine
New Haven CT 06519
USA, and
Howard Hughes Medical Institute
Chevy Chase, MD 20815-6789
USA
Tel. 1-203-392-4167
Email: tolulope.agunbiade@yale.edu

Preface

Insects are the most successful group of animals on the planet and are ecologically and economically extremely important. Today's entomology field has gone beyond borders and is termed as a "super science." With its multidisciplinary approach, entomology explores new scientific frontiers. It has emerged to provide some of the most powerful tools in resolving fundamental biological questions and problems using genomic (genome sequencing, assigning functions to genes, determining genome architecture) and proteomic (nature of proteins, 3D structure, posttranscriptional modifications) approaches.

Whole-genome sequence projects for insect model organisms (29 insect species completed and many more under way) and the concurrent growth of sequence databases provide the biological sciences with invaluable sources of information. The two volumes of this book are intended to share the efforts of major contributors with the genomic and proteomic communities. This will pave the way toward the development of new and innovative approaches to improve public health and agriculture using effective and ecologically sound pest management systems.

We therefore decided to create an up-to-date reference that would provide a firm basis for understanding the past and current genomic and proteomic research conducted in entomology. To do this, we decided to bring together leading world scientists in molecular entomology and biotechnology to share their past experiences in the development of this field, to summarize the current state of the art, and to offer hypotheses and predictions to set a framework for future research. This book is composed of volumes 1 and 2 with 18 chapters.

Volume 1 *Short Views on Insect Genomics*

The first volume presents 8 chapters that address genomic approaches currently employed using various model organisms: body lice, whitefly, aphid, *Drosophila*, mosquitoes, lepidopterans, and others.

Chapter 1 provides a detailed story of the body louse genome project and its significance in understanding fundamental questions about the biology of lice and their endosymbionts. Of importance, it discusses the use of reverse genetics (RNAi) to answer questions about the role of specific genes in biological processes (Barry R. Pittendrigh, University of Illinois, USA).

Chapter 2 addresses advances in the genomics of the whitefly, *Bemisia tabaci*, an insect pest and plant virus vector. It also discusses the interaction between insects and viruses and the development of control strategies using RNAi approaches (Murad Ghanim, Agricultural Research Organization, Israel).

Chapter 3 provides an update on the scope and scale of the genomic data of Lepidoptera that is available in public databases and discusses the current status of lepidopteran genome projects. Special attention is drawn to (i) *Elongation factor-1α*, (ii) *Wingless*, (iii) *Cytochrome c oxidase I*, and (iv) *ribosomal DNA and RNA* (América Nitxin Castañeda-Sortibrán, Universidad Nacional Autónoma de México).

Chapter 4 deals with the genetic and molecular mechanisms underlying the evolution of different aphid biotypes with respect to naturally occurring host plant resistance (Andy Michel, Ohio State University, USA).

Chapter 5 presents integrative genomic approaches used in studying epigenetic mechanisms of phenotypic plasticity in the pea aphid. It describes how epigenetic mechanisms (DNA methylation and chromatin remodeling) play an increasingly important role in winged vs wingless polyphenism in this highly adaptable species (Gael LeTrionnaire, Institut National de la Recherche Agronomique, France).

Chapter 6 provides an introduction to the concepts behind the dynamic and powerful field of insect regulatory genomics. It describes successful strategies and techniques for finding regulatory elements in model insect species like *Drosophila*, current efforts to extend them to evolutionarily diverged non-model organisms, and potential applications of this information using such approaches as gene transfer and RNAi (Marc S. Halfon, University at Buffalo-State University of New York, USA).

Chapter 7 presents a comprehensive coverage of comparative genomics of transcription factor binding in *Drosophila* by using ChIP-Chip, ChIP-Seq, and DamID techniques to discover a deeper understanding of genomic regulatory mechanisms (Steven Russell, University of Cambridge, UK).

Without the application of bioinformatics, the growth of genomic and proteomics would be limited. Chapter 8 focuses on a machine learning approach (ClanTox, NeuroPID, TOLIPs) to discover short bioactive proteins and peptides from insect genomes (Michal Linial, the Hebrew University of Jerusalem, Israel).

Volume 2 *Short Views on Insect Proteomics*

The second volume presents comprehensive and cutting-edge studies with emphasis on proteomics. It comprises 10 chapters which constitute a key reference manual for everyone involved in insect biochemistry, molecular genetics, molecular evolution,

insect bioinformatics and structural biology, applications of insect biotechnology, insect "omics," and related fields.

Ticks transmit viral diseases to livestock, which are of great economic importance worldwide. Chapter 1 focuses on ticks (blood-sucking parasite) and recent developments in the field of sialomes (salivary gland proteomes). It discusses the regulation of host hemostasis and the molecular immune mechanisms behind it. It also discusses the utilization of salivary gland proteins in vaccines to control vector-borne diseases (Youmna M'ghirbi, University of Tunis El-Manar, Tunisia).

Current proteomic approaches rely on the application of mass spectrometry to protein molecules. Chapter 2 describes qualitative and quantitative proteomic methods for the analysis of the *Anopheles gambiae* mosquito proteome with emphasis on circadian changes in expression (Giles E. Duffield, University of Notre Dame, USA).

Chapter 3 reviews recent advances in the knowledge of the lepidopteran digestive system. Key topics include the architecture, structure, and function of the lepidopteran peritrophic matrix (Dwayne D. Hegedus, Agriculture and Agri-Food Canada, Canada).

Many key agents protect insects from injury at low temperatures. Chapter 4 documents cold adaptation responses in insects and other arthropods using an "omics" approach (Duško P. Blagojević, University of Belgrade, Serbia).

Chapter 5 presents evidence for the evolutionary extinction of enzyme and molecular systems that engage and utilize the nonstandard amino acid, selenocysteine, in insects (Marco Mariotti, Centre de Regulació Genòmica, Barcelona, Spain).

Chapter 6 highlights recent progress in understanding the mechanisms behind the insect innate immune response with the silkworm, *Bombyx mori*, as a model organism. It reviews the characteristic features of antibacterial proteins and antimicrobial peptides (AMPs) produced by insects against pathogens, their modes of action, and current and potential medical applications of these molecules (Chandan Badapanda, Xcelris Genomic Research Center, India).

Chapter 7 takes the reader to the post genomic era where insects have become important models for applied sciences. This chapter describes the use of insect cell lines derived from model organisms like *Bombyx mori* as expression systems for vaccines and other peptides and proteins, and the use of advanced protein expression systems based on the *B. mori* nucleopolyhedrovirus (BmNPV) bacmid (Enoch Y. Park, Shizuoka University, Japan).

Chapter 8 concentrates on the use of insects and their associated microorganisms as an important resource in diverse industries, especially for the production of industrial enzymes, microbial insecticides, and many other substances (Anthony Ejiofor, Tennessee State University, USA).

Chapter 9 deals with the special structure and properties of spider silks and their biotechnological applications (Daniela Matias de C. Bittencourt, Brazilian Agricultural Research Corporation, Brazil).

Chapter 10 focuses on the development, properties, and application of nanoparticles derived from plants producing bioactive compounds for use as novel agents to control human and insect pests (K. Murugan, Bharathiyar University, India).

It is our pleasure to launch the twin volumes of *Short Views on Insect Genomic and Proteomics*, in the Springer series, 3 & 4. The reader will find a wide variety of topics addressed in detail, which will help them update their knowledge of insect genomics and proteomics.

Manhattan, KS, USA
Kingston, RI, USA
New Haven, CT, USA
Chevy Chase, MD, USA

Chandrasekar Raman
Marian R. Goldsmith
Tolulope A. Agunbiade

Acknowledgement

I wish to express my thanks and gratitude to authors from all over the world for contributing their outstanding chapters (USA, UK, Canada, Mexico, Israel, Brazil, Serbia, Tunisia, Germany, Spain, France, Ireland, Turkey, South Korea, China, Japan, India). It was a pleasure working with this expert team of scientists, and I would gladly do so again in a moment's notice. We will miss this collaboration now that it has ended, but will feel rewarded if this book is appreciated by our team/colleagues in the field of entomology.

The chapters of this book series (Volumes 1 and 2) are organized to present many experts' contributions – highlighting their current lab research – to provide an overview of current and prominent advances in insect genomics and proteomics, and biotechnology, which will help students and researchers to broaden their knowledge and to gain an understanding of both the challenges and the opportunities behind each approach.

We collectively called the volumes *Short Views on Insect Genomics and Proteomics* with the support of Springer Science Media. I would like to extend my sincere gratitude to Prof. Cônsoli L. Fernando for approving this project. It was a really wonderful opportunity to work with the Springer editorial team (Dr. Zuzana Bemhart, Senior Publishing Editor; Mariska van der Stigchel, Editorial Assistant; Dr. William F. Curtis, Executive Vice President; Dr. Jacco Fllipsen, Vice President; Dr. Sadie Forrester, Executive Editor) and many more from the Springer family. We apologize to those whose work could not be cited owing to space considerations as well as any errors that may have occurred in the text, and we would be grateful to receive any comments or suggestions about improvements for further editions.

I thank my associate editors (Prof. Marian Goldsmith, University of Rhode Island, USA, and Dr. Tolulope A. Agunbiade, Yale University School of Medicine, USA, and Howard Hughes Medical Institute, USA) for their continuous support and vigilance over the book project and for always giving advance notice of the editing and proofreading schedules. Most importantly, I want to express appreciation to my wife (P.G. Brintha, "Division of Biology", Kansas State University, USA), who in all possible ways helped to transform our original efforts into an acceptable final form.

I also thank my Prof. Gerald Reeck, Department of Biochemistry and Molecular Biophysics, Kansas State University, for providing encouragement and support throughout.

I also wish to take this opportunity to express my gratitude to my former teacher, Prof. Seo Sook Jae, (GSNU, South Korea), Prof. Yeon Soo Han, (Chonnam National University, South Korea), Prof. M. Krishnan (Bharathidasan University, India), Prof. Subba Reddy Palli (University of Kentucky, USA), and other external mentors, Prof. Enoch Y. Park (Shizuoka University, Japan), Prof. M. Kobayashi (Nagoya University, Japan), Dr. Hiroaki Abe (Tokyo University of Agri. and Technology, Japan), Prof. M. Takeda (Kobe University, Japan), Dr. A.P.J. Abdul Kalam (Ex-President, India), Prof. Jang Hann Chu (National University of Singapore, Singapore), Prof. Ji-Ping Liu (Asia-Pacific Sericulture Training Center, China), Prof. Thomas W. Sappington (USDA-ARS,USA), Prof. Fernando G. Noriega (Florida International University, USA), Dr. Srinivasan Ramasamy (AVRDC, The World Vegetable Center, Taiwan), Dr. B.K. Tyagi (ICMR, India), Dr. H.C. Sharama (ICRISAT, India), Prof. Lawrence I. Gilbert (North Carolina University, USA), Prof. Subbaratnam Muthukrishnan (Kansas State University, USA), Prof. Manickam Sugumaran (University of Massachusetts, USA), Prof. František Sehnal (Institute of Entomology, Czech Republic), Prof. Anthony A. James (University of California, USA), Prof. David O'Brochta (University of Maryland, USA), Prof. Xavier Belles (CSIC-UPF, Spain), Dr. Emmanuelle Jacquin-Joly (INRA UPMC, France), Prof. Immo A. Hansen (New Mexico State University, USA), and Prof. Lizette Koekemoer (University of the Witwatersrand, South Africa), whose inspiration supported me in many ways for the commencement of this "International Book Mission Program."

The seed for this International Book Mission project was planted and initially nurtured by Prof. Rolando Rivera-Pomar (Universidad Nacional del Noroeste de Buenos, Argentina). The book mission project was initiated in April 2013, completed in July 2015, and published in December 2015.

Kansas State University
Manhattan, KS, USA

Chandrasekar Raman

Contents

Contents of Volume 1

Contents of Volume 2

Contributors

Volume 1

Raman Bansal Department of Entomology, Ohio Agricultural Research and Development Center, The Ohio State University, Wooster, OH, USA

Jennifer A. Brisson Department of Biology, University of Rochester, Rochester, NY, USA

Sarah Carl Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK

América Nitxin Castañeda-Sortibrán Departamento de Biología Celular, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de México, México

J.M. Clark Department of Veterinary and Animal Science, University of Massachusetts, Amherst, MA, USA

Murad Ghanim Department of Entomology, Volcani Center, Bet Dagan, Israel

Mary Grantham Department of Biology, University of Rochester, Rochester, NY, USA

Marc S. Halfon Department of Biochemistry, University at Buffalo-State University of New York, Buffalo, NY, USA

Department of Biological Sciences, University at Buffalo-State University of New York, Buffalo, NY, USA

Department of Biomedical Informatics, University at Buffalo-State University of New York, Buffalo, NY, USA

Program in Genetics, Genomics and Bioinformatics, University at Buffalo-State University of New York, Buffalo, NY, USA

NY State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY, USA

Molecular and Cellular Biology Department and Program in Cancer Genetics, Roswell Park Cancer Institute, Buffalo, NY, USA

Surapathrudu Kanakala Department of Entomology, Volcani Center, Bet Dagan, Israel

Gael Le Trionnaire UMR 1349 (INRA – Agrocampuus Ouest – University of Rennes I) IGEPP – Institute of Genetics Environment and Plant Protection, Rennes, Le Rheu cedex, France

S.H. Lee Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul, Republic of Korea

Department of Agricultural Biotechnology, Seoul National University, Seoul, Republic of Korea

Michal Linial Department of Biological Chemistry, Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel

Andy Michel Department of Entomology, Ohio Agricultural Research and Development Center, The Ohio State University, Wooster, OH, USA

Dan Ofer Department of Biological Chemistry, Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel

B.R. Pittendrigh Department of Entomology, University of Illinois, Urbana/Champaign, IL, USA

Carmen Pozo Departamento de Conservación de la Biodiversidad, El Colegio de la Frontera Sur, Chetumal, Quintana Roo, México

Blanca Prado Departamento de Conservación de la Biodiversidad, El Colegio de la Frontera Sur, Chetumal, Quintana Roo, México

Nadav Rappoport School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel

Steven Russell Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Cambridge, UK

K.M. Seong Department of Entomology, University of Illinois, Urbana/Champaign, IL, USA

L.D. Steele Department of Entomology, University of Illinois, Urbana/Champaign, IL, USA

W. Sun Department of Entomology, University of Illinois, Urbana/Champaign, IL, USA

Kushal Suryamohan Department of Biochemistry, University at Buffalo-State University of New York, Buffalo, NY, USA

NY State Center of Excellence in Bioinformatics and Life Sciences, Buffalo, NY, USA

Denis Tagu UMR 1349 (INRA – Agrocampuus Ouest – University of Rennes I) IGEPP – Institute of Genetics Environment and Plant Protection, Rennes, Le Rheu cedex, France

K.S. Yoon Department of Biological Sciences and Environmental Sciences Program, Southern Illinois University-Edwardsville, Edwardsville, IL, USA

Volume 2

Mohamad Saleh AlSalhi Department of Physics and Astronomy, King Saud University, Riyadh, Kingdom of Saudi Arabia

Chandan Badapanda Xcelris Genomics Research Centre, Ahmedabad, India

OPJS University, Churu, Rajasthan, India

Daniela Matias de C. Bittencourt Research and Development Department, Brazilian Agricultural Research Corporation – Embrapa, Brasília, Brazil

Duško P. Blagojević Department for Physiology, Institute for Biological Research, University of Belgrade, Belgrade, Republic of Serbia

M.M. Champion Department of Chemistry and Biochemistry, Nieuwland Science Hall, University of Notre Dame, Notre Dame, IN, USA

Surendra K. Chikara Xcelris Genomics Research Centre, Ahmedabad, India

S. Devanesan Department of Physics and Astronomy, King Saud University, Riyadh, Kingdom of Saudi Arabia

D. Dinesh Department of Zoology, Bharathiar University, Coimbatore, India

Giles E. Duffield Department of Biological Sciences, Galvin Life Science Center, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

Anthony Ejiofor Department of Biological Sciences, College of Agriculture, Human and Natural Sciences, Tennessee State University, Nashville, TN, USA

Martin Erlandson Agriculture and Agri-Food Canada, Saskatoon, SK, Canada

Department of Biology, University of Saskatchewan, Saskatoon, SK, Canada

S.A. Freed Department of Biological Sciences, Galvin Life Science Center, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

Gordana Grubor-Lajšić Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Republic of Serbia

Dwayne D. Hegedus Agriculture and Agri-Food Canada, Saskatoon, SK, Canada

Department of Food and Bioproduct Sciences, University of Saskatchewan, Saskatoon, SK, Canada

Jiang-Shiou Hwang Institute of Marine Biology, National Taiwan Ocean University, Keelung, Taiwan

Tatsuya Kato Green Chemistry Research Division, Research Institute of Science and Technology, Shizuoka University, Shizuoka, Japan

Danijela Kojić Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Republic of Serbia

Youmna M'ghirbi Laboratory of Veterinary Microbiology and Epidemiology, Service of Medical Entomology, Institute Pasteur of Tunis, University of Tunis El-Manar, Tunis, Tunisia

P. Madhiyazhagan Department of Zoology, Bharathiar University, Coimbatore, India

Marco Mariotti Bioinformatics and Genomics Programme, Centre for Genomic Regulation, Barcelona, Catalonia, Spain

Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain

School of Biochemistry and Cell Biology, University College Cork, Cork, Ireland

Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

K. Murugan Department of Zoology, Bharathiar University, Coimbatore, India

J.E. O'Tousa Department of Biological Sciences, Galvin Life Science Center, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

C. Panneerselvam Department of Zoology, Bharathiar University, Coimbatore, India

Enoch Y. Park Green Chemistry Research Division, Research Institute of Science and Technology, Shizuoka University, Shizuoka, Japan

Edward Petri Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Republic of Serbia

Željko D. Popović Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Republic of Serbia

Jelena Purać Department of Biology and Ecology, Faculty of Sciences, University of Novi Sad, Novi Sad, Republic of Serbia

Chandrasekar Raman Department of Biochemistry and Molecular Biophysics, Kansas State University, Manhattan, KS, USA

S.S.C. Rund Department of Biological Sciences, Galvin Life Science Center, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

A.D. Sheppard Department of Biological Sciences, Galvin Life Science Center, Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA

J. Subramanium Department of Zoology, Bharathiar University, Coimbatore, India

Umut Toprak Department of Plant Protection, College of Agriculture, University of Ankara, Ankara, Turkey

Jiang Wei Department of Microbiology, College of Biological Sciences, China Agricultural University, Beijing, China

Chapter 1
Body Lice: From the Genome Project to Functional Genomics and Reverse Genetics

B.R. Pittendrigh, J.M. Clark, S.H. Lee, K.S. Yoon, W. Sun, L.D. Steele, and K.M. Seong

Abstract In 2010, the initial annotations of the genomes of the body louse (*Pediculus humanus humanus* Linnaeus) and its primary endosymbiont, "Candidatus *Riesia pediculicola*," were completed. The body louse had the smallest genome of any insect sequenced to that point. Prior to the proposal for the sequencing project, there was a dearth of information about louse genes, with no more than around 500–600 inferred open reading frames in public databases. Since the publishing of this genome project, the field of louse genomics has experienced significant advances in our understanding of the taxonomic relationship and the differences in vector competence between head and body lice. To date, the louse system has emerged as a model system to understand xenobiotic induction responses. Finally, a louse RNAi-based reverse genetic system has been developed with the potential to study the functional role of louse genes in vector competence.

B.R. Pittendrigh (✉) • W. Sun • L.D. Steele • K.M. Seong
Department of Entomology, University of Illinois, Urbana/Champaign, Urbana, IL 61801, USA
e-mail: pittendr@illinois.edu; wsn@illinois.edu; steele11@illinois.edu; kseong6@illinois.edu

J.M. Clark
Department of Veterinary and Animal Science, University of Massachusetts, Amherst, MA 01003, USA
e-mail: jclark@vasci.umass.edu

S.H. Lee
Research Institute of Agriculture and Life Sciences, Seoul National University, Seoul 151-921, Republic of Korea

Department of Agricultural Biotechnology, Seoul National University, Seoul 151-921, Republic of Korea
e-mail: shlee22@snu.ac.kr

K.S. Yoon
Department of Biological Sciences and Environmental Sciences Program, Southern Illinois University-Edwardsville, Edwardsville, IL 62026, USA
e-mail: kyoon@siue.edu

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*, Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_1

Abbreviations

ABC	ATP-binding cassette
BAC	Bacterial artificial chromosome
dsRNA	Double-stranded RNA
Est	Esterase
GST	Glutathione-S-transferase
JCVI	J. Craig Venter Institute
MCE	Malathion carboxylesterase
mtSSB	Mitochondrial single-stranded binding protein
NCBI	National Center for Biotechnology Information
NIH	National Institutes of Health
OBP	Odorant-binding protein
OR	Odorant receptors
ORF	Open reading frame
P450	Cytochrome P450
RNAi	RNA interference
TE	Transposable element

1.1 Introduction and Background on the Need for a Louse Genome Project

The late 1990s and early 2000s marked a transition in insect molecular biology away from traditional work focused on single genes to the genomics era when insect genomes and functional genomics showed the promise of transforming our understanding of insect systems. As late as 2005, however, only a single analysis of open reading frames (ORFs) of the body louse transcriptome, encompassing 506 inferred ORFs, had been published [1], along with a handful of other NCBI (National Center for Biotechnology Information) entries. This lack of genomic knowledge of lice stood in stark contrast to their potential to reveal critical information on important/fundamental biological phenomena. Outlined below are several of the key and interesting aspects of louse biology, which served to gain from the entry of this model system into the genomics era.

Carl Linnaeus first described the body louse (*Pediculus humanus humanus* Linnaeus) in 1758, and less than a decade later (in 1767), it was differentiated from the head louse (*Pediculus humanus capitis* de Geer). However, even two centuries later, their status as separate species was still being debated in the literature [2, 3]. Although head and body lice do not interbreed in the wild [4, 5], fertile first-generation (F1) hybrids have been produced under laboratory conditions [6, 7]. Importantly, head and body lice have specialized into infesting specific regions on their human hosts. Head lice live exclusively in the scalp region of humans where

the females attach their eggs to the base of hair shafts. In contrast, body lice feed in the non-head regions of humans and the females secure eggs to clothing [8]. Head and body lice have classically not been thought to migrate between the head and body regions even when dual infestations have occurred. Head and body lice also differ in their morphologies. Body lice are larger than head lice (Fig. 1.1) and feed less often; body lice are thought to be larger as a result of larger but less frequent blood meals [8, 9]. Thus, head and body lice represent an excellent example of two closely related groups that have adopted separate ecological niches; however, unlike many other examples of this phenomenon, their food supply remains the same – human blood. From a genomics perspective, this makes them an interesting system to understand the molecular and evolutionary implications of populations living in two different ecological niches – the head and the body.

Although head and body lice also differ in their capacity to vector three important human diseases, both have made significant social impacts at different points in human history. In the past several decades, head lice have become a concern from a medical perspective, with body lice receiving far less attention. Head lice can be commonly found in school-aged children with symptoms that include itching and in some cases loss of sleep [8, 10] and loss of school time and work time for their parents. Despite posing considerable inconvenience/annoyance, head lice are not known to be vectors of human diseases. In contrast, body lice are competent vectors of three important human diseases [8] and are known to transmit three different

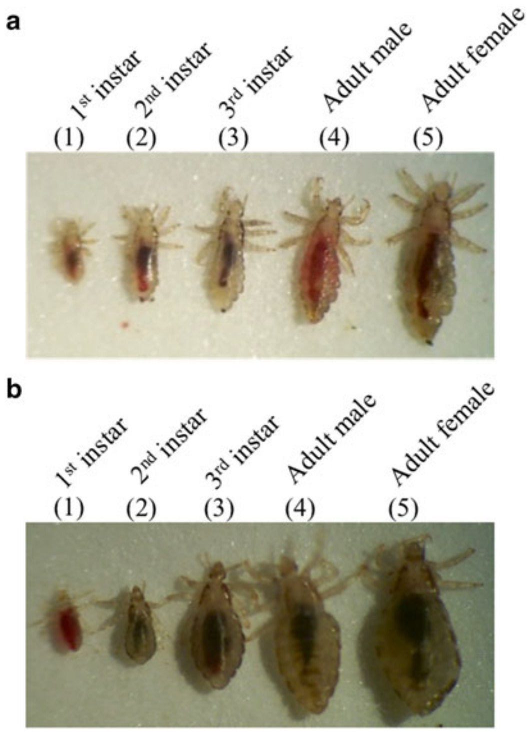


Fig. 1.1 Developmental stages of (**a**) head lice and (**b**) body lice, including (*1*) first instar, (*2*) second instar, (*3*) third instar, (*4*) adult males, and (*5*) adult females

species of bacteria (*Bartonella quintana*, *Borrelia recurrentis*, and *Rickettsia prowazekii*) to their human hosts [8, 9]. *B. quintana* is the causative agent of trench fever, an extremely important disease during the World War I conflict which was thought to impact approximately one million people, mainly soldiers. This disease still remains a problem in areas of the world where human living conditions have been impacted by man-made or natural disasters [7, 9, 11]. *B. recurrentis* causes a disease in humans called relapsing fever, which, if left untreated, can result in mortality rates in humans up to 70 %. Although of greater medical relevance in previous decades and centuries, this disease still remains a significant problem in regions of the world where living conditions are poor and where people have limited access to medical care [12]. Finally, during World War I, epidemic typhus, caused by *R. prowazekii*, resulted in an excess of three million deaths as a result of crowded prison conditions. Body lice are seldom observed in the USA and much of the rest of the developed world. They do remain, however, a public health threat due to infestations that occur in homeless individuals or those that have been affected by natural or man-made disasters (e.g., the 2008 financial crisis) that lead to reduced hygienic conditions [11].

Other noteworthy biological aspects/features of head and body lice made the study of their genomes of interest to the scientific community. Both head and body lice actively maintain an obligate endosymbiont, "Candidatus *Riesia pediculicola*," and the genomic and biochemical relationship between lice and the endosymbiont was greatly facilitated by the sequencing of their respective genomes. Additionally, both humans and chimpanzees maintain louse populations that are of common origin, just like their mammalian hosts. The fact that these obligate ectoparasites have maintained such a long coexistence with these closely related hosts has the potential to enable the scientific community to ask questions about the rates and patterns of evolutionary change in host versus parasite populations.

1.2 Events Leading to the Body Louse Genome Project

As of 2005, the louse community did not have the genomics tools available to it to make significant headway toward understanding the molecular components of the aforementioned issues of human louse biology. Several events transpired resulting in an opportunity that would ultimately allow for the sequencing and annotation of the genome and a community effort to describe the biological significance of the genome. Prior to 2005, the Pittendrigh, Clark, and Lee laboratories (coauthors on this chapter) had been working together on toxicological issues associated with louse biology. Then, in 2005, an opportunity came about to have the genome size of the body louse tested (by Dr. Spencer Johnston, Texas A&M University) using a louse colony maintained in the Clark laboratory. Dr. Johnston observed that the body louse had the smallest genome of any insect he had measured to that point, an observation later published in Johnston et al. [13]. This provided the impetus for a small group of individuals in the louse and insect genomics community to submit a

white paper to the National Institutes of Health (NIH) to seek funding for sequencing the body louse genome [14, 15]. The sequencing of the genome was ultimately funded by NIH and took place at the J. Craig Venter Institute (JCVI) with Dr. Ewen Kirkness leading the sequencing, and they worked with VectorBase on the annotation efforts. Although the CG content of the body louse genome did not allow for the effective creation of a bacterial artificial chromosome (BAC) library to facilitate genome construction and analysis, the small size and very limited number of repetitive elements allowed for an assembly using data from shotgun sequencing. Analysis of the genome took place over several years through a virtual online community that met and collaborated through conference calls, Skype calls, Google Docs, and e-mails. The genome was published in 2010 [16] and included an analysis of its major aspects. Two supporting papers that focused more extensively on the louse detoxification system [17] and the louse RNAi machinery followed soon afterward [18].

1.3 Initial Outcomes of the Body Louse Genome Sequencing Project

In addition to an ab initio prediction of the body louse genome [16], its primary bacterial endosymbiont, "Ca. *R. pediculicola*", was also sequenced and described. The body louse genome itself contained a relatively small number of genes for an insect species, with 10,773 predicted protein-coding genes (Table 1.1). A very high

Table 1.1 Summary of the genome features of *Pediculus humanus humanus* compared with *Drosophila melanogaster*

Genome feature	Count	Nucleotides (Mb)	Genome fraction (%)
P. h. humanus (*D. melanogaster*)	6 chromosomes (4 chromosomes)	110 (169)	100 (100)
Gene-rich clusters containing 95 % of genes	1110 (1130)	55 (70)	50 (41)
Protein-coding genes			
Total [multi-exon]	10,773 [10,424]; (13,794 [11,458])	33.8 (82.6)	31 (49)
Coding exons	69,261 (54,606)	16.6 (22.3)	15 (13)
Introns	58,522 (44,698)	17.2 (48.6)	15 (29)
Nonprotein-coding genes			
tRNAs	161 (292)	0.012 (0.022)	<1
miRNAs	57 (90)	0.005 (0.008)	<1
Transposable elements	3558 (9409)	1.1 (11.6)	1 (7)
Tandem repeats	130,608 (25,904)	6.9 (6.1)	6 (4)

Parentheses, *D. melanogaster*
Originally published by Kirkness et al. [16] and reprinted with permission

percentage of these genes had known orthologues with other species, with only 163 of them having no known orthologues. Additionally, 57 microRNA genes were predicted. A follow-up transcriptome analysis of both the head and body louse was published by Olds et al. [19] which demonstrated that, notably, in the original whole genome predictions by Kirkness et al. [16], only a single microRNA had been missed. Additionally, the Olds et al. [19] analysis of the head louse transcriptome revealed that head and body lice had almost the exact same numbers of genes, but failed to find one gene in the head louse (PHUM540560) known to be present in body lice. Drali et al. [20], in a subsequent analysis, observed that this gene did occur in head lice; however, PHUM540560 had significant sequence differences between head and body lice. The combined findings of Olds et al. [19], Kirkness et al. [16], and Drali et al. [20] provide strong evidence that the annotation of the body and head louse genomes likely represents a complete analysis of the microRNA and protein-coding genes in these insects. Subsequently, whole genome sequences of head lice were determined by next-generation sequencing methods and compared with the reference genome sequences of body lice (SH Lee, unpublished data). The total consensus showed a size of 110 Mbp and 96 % coverage of body louse genome sequences. A single nucleotide polymorphism analysis revealed a nucleotide diversity of 2.2 % between body and head lice genomes, which was larger than that of transcriptome differences between body and head lice. Thus, the louse community is now well-positioned to ask questions of louse biology using functional and structural genomic approaches, as well as through comparative evolutionary analyses.

One of the significant challenges of describing a genome is to place it within the context of the life history of the organism. Although the body louse is a parasite, with a reduced need for many systems that a free-living nonparasitic organism would need (e.g., for senses and responding to diverse environmental factors), it is an ectoparasite that is still somewhat independent from its host. What emerged from the genome project was very much in keeping with what we know about louse biology. First, the body louse contains a relatively complete set of genes associated with basic cellular processes, something that one would expect for an ectoparasite. However, it is an ectoparasite that lives, essentially, in a very simple "ecological niche"; the body louse only has to sense the host and feed on blood. Thus, one would predict a reduction in the number of genes associated with the body louse's ability to sense and respond to its environment. This assumption was, in fact, the case, as the louse genome has a reduced number of genes compared to, for example, *Drosophila melanogaster*, in pathways associated with these processes. Categories of genes with reduced numbers include (i) G protein-coupled receptors; (ii–iv) odorant-, gustatory-, and chemosensory-related genes, and (v–viii) cytochrome P450s (P450s); glutathione-S-transferases (GSTs); and esterases (Ests) [16]. For example, the body louse genome contains only 37 P450s, the smallest number observed in any insect system [21]. With such a limited number of P450s, human body lice represent an excellent system for the study of inducible P450s that have been further investigated for their role in pesticide tolerance and perhaps resistance [22].

Another major theme to emerge out of the genome project was that the small size of the body louse genome was at least partially due to a very low level of

transposable elements (TEs), about 1 % of the genome. This was lower than the level of TEs of any other insect genome that had been sequenced to that date [16]. It is important to note that the genome size of both body and head lice are at the critical threshold where it is considered that TEs can become established in eukaryotic genomes [16]. Thus, the genome size itself may play a role in the small number of TEs.

The genome project also resulted in the sequencing of the body louse mitochondrial genome. This structure/feature is of particular interest as, instead of the usual single large self-replicating circular chromosome, the body louse has 18 minicircular mitochondrial chromosomes. Although the mitochondrial genome was relatively complete in terms of its expected genetic composition, it was missing the mitochondrial single-stranded DNA-binding protein (mtSSB). The lack of mtSSB is thought to explain the minicircular nature of the mitochondrial chromosomes, as this protein is involved in the optimal initiation and processivity during critical molecular steps associated with mitochondrial genome replication [16].

Finally, during the sequencing of the body louse genome, as an added bonus, the obligatory louse endosymbiont, "Ca. *R. pediculicola*," was also sequenced [16]. The genome of this endosymbiont contained less than 600 genes, which were arranged either on a circular plasmid or on a second short, linear chromosome. Lice are known to rely on this endosymbiont for pantothenate in order to survive – centrifugation of eggs to dislodge the egg mycetomes (which house the endosymbionts) kills most of the endosymbionts and has serious negative consequences on louse biology [23]. The circular plasmid from the "Ca. *R. pediculicola*" genome possesses an arrangement of genes associated with the synthesis of pantothenate. The louse genome is devoid of these genes, and presumably, pantothenate is not sufficient in the diet of lice [16]. Of potential practical application is the fact that the bacterial genome does not contain any antibiotic resistance genes [16], suggesting the potential for the development of antibiotic-based louse control strategies. Although this is theoretically possible, the practical issue of (1) how such antibiotics could be delivered to the lice, (2) whether they would be effective in controlling louse populations, and (3) the regulatory, legal, and risk issues associated with this type of antibiotic use remain to be addressed.

1.4 The Body Louse Genome as a Foundation for Asking Important Biological Questions

Since its publication, the body louse genome has served as a reference genome for other insect genome projects and gene family studies. It has been used in comparative genomic research with other important insects such as the bed bug (*Cimex lectularius*) [24], the fire ant (*Solenopsis invicta*) [25], psocids (*Liposcelis entomophila*) [26], leaf-cutter ants (*Atta cephalotes*) [27], and diamondback moth (*Plutella xylostella*) [28]. The comparison with these other insect genomes provided a foundation for future functional genomics studies. In addition, body louse genome

research has provided insights into important biological processes such as olfaction [29], insect immunity [30], gene regulation [31], and evolution of resistance to xenobiotics in other insect systems [32]. Because of the comparatively small number of genes in the body louse genome, it was used as a reference in the latter studies [32, 33]. For example, this provided the ability to understand which P450, EST, and GST genes may play a role in resistance to xenobiotics. The body louse genome has also been used in analyses for the broader understanding of olfactory evolution, neurobiology, and sensory processing, including the gene families encoding odorant receptors (ORs) and odorant-binding protein (OBP) [29, 34, 35]. Thus, the body louse genome will continue to be used as a reference for other insect genomic studies.

The analysis of the body louse genome has provided the louse community with the initial tools to begin to ask fundamental questions regarding basic issues of louse biology. First, what are the structural genomic differences, if any, between body and head lice? Second, are there significant transcriptional differences between head and body lice? Third, do body and head lice respond differently, at a transcriptional level, to microbial agents, especially diseases that body lice are known to transmit to humans? Fourth, can such transcriptional differences be used to develop a hypothesis as to why body lice transmit human disease agents and head lice do not? The body louse genome has also allowed the scientific community to address other broad biological questions in the areas of evolutionary biology and toxicology [16–21].

1.5 Vector Competence

To date, two studies have provided initial insights regarding head and body louse responses to bacterial infections – one with Gram-positive and Gram-negative bacteria [36] and a second using *B. quintana* [37]. As head and body lice have nearly identical genomes, and different vector competences, they have thus provided an interesting model system for the study of vector competence.

Kim et al. [36] investigated the differences in immune response between head and body lice following injections with one of two model bacteria, *Staphylococcus aureus* (Gram-positive) or *Escherichia coli* (Gram-negative), into the lice and measured the proliferation rates of these bacteria postinjection (Fig. 1.2). Body lice exhibited a reduced immune response compared to the head lice; this difference was more acute with *E. coli*, especially at the initial stages of infection. Although the body louse genome has fewer immune-related genes and lacks a functional immune deficiency (Imd) pathway compared to other insects, it still has all the required components of the other major insect immune pathways. Kim et al. [36] also determined a transcriptional profile of the genes involved in the louse humoral immune response. Both head and body lice had an increased immune response to *S. aureus*, but a minimal response was mounted after *E. coli* treatment (Fig. 1.3). Additionally,

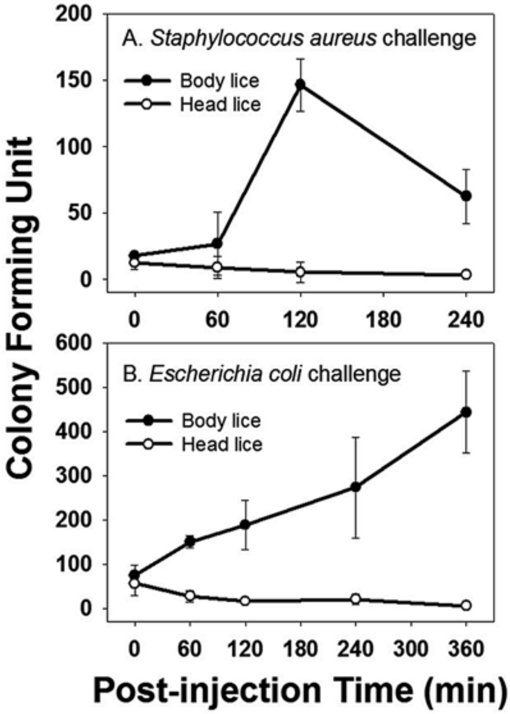


Fig. 1.2 Time course of bacteria proliferation (colony-forming unit) inside body and head lice injected with equal concentrations of *Staphylococcus aureus* (**a**) and *Escherichia coli* (**b**) (Figure and figure legend are reproduced from Kim et al. [36] with permission from Elsevier)

head lice had a greater phagocytotic (cellular) response to *E. coli* compared with body lice (Fig. 1.4). In contrast, head and body lice displayed only slight differences in their phagocytotic activity against *S. aureus*. Their observations support the hypothesis that increased vector competence in body lice relative to head lice may be due to a reduced humoral and cellular immune response in body lice.

Previte et al. [37] used an in vitro louse-rearing system in order to infect head and body lice with human blood containing *B. quintana.* Subsequently, they investigated the differences in (1) the proliferation of *B. quintana* and (2) transcriptional differences of immune-related genes in head and body lice. *B. quintana* proliferation was greater in body lice than in head lice, showing significantly higher levels in body lice 6–12 days postinfection (Fig. 1.5). Transcriptional comparisons revealed differences between head and body lice in eight immune response genes. Many of these genes are known to be associated with the Toll pathway: defensin 1, spaetzle, serpin, fibrinogen-like protein, apolipophorin 2, and scavenger receptor A. Their findings supported the hypothesis that head lice fight the *B. quintana* infection earlier than do body lice, in keeping with the hypothesis that a "poorer" immune system in body lice may be the basis of their ability to vector *B. quintana*. It remains to be determined if this is the underlying basis for the ability of both lice to vector other human diseases.

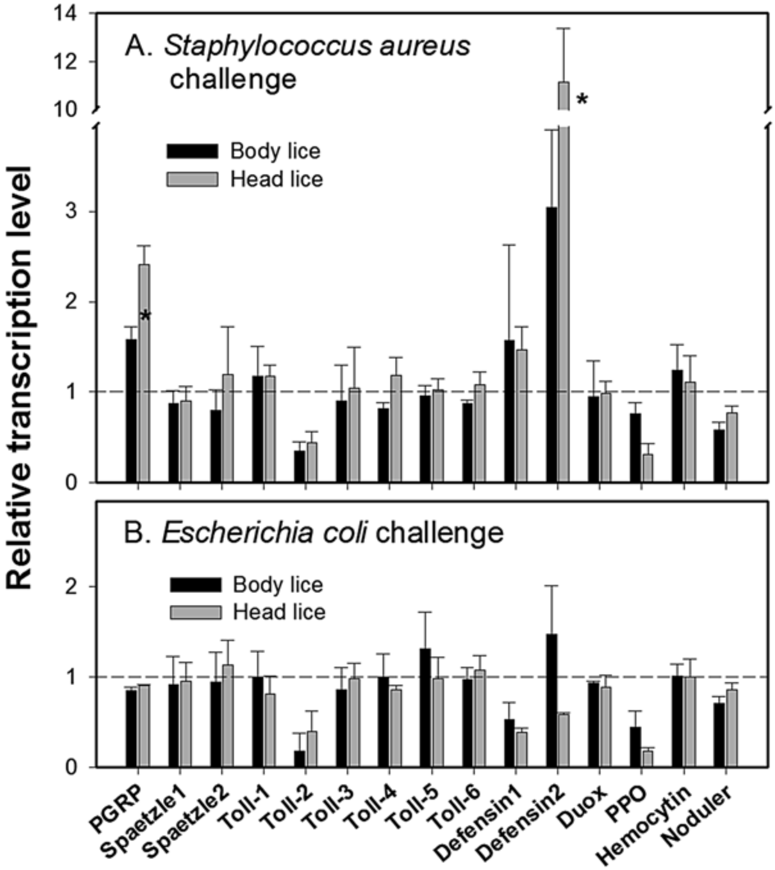


Fig. 1.3 Comparison of the relative transcription level of 15 representative immune-related genes following challenges by *Staphylococcus aureus* (**a**) and *Escherichia coli* (**b**) in the adult stage of body and head lice. *Bars* with *star mark* (*) indicate statistical difference between body and head lice ($p < 0.05$) (Figure and figure legend are reproduced from Kim et al. [36] with permission from Elsevier)

1.6 Host-Parasite Evolution

The body louse genome provided the basis for a recent sequencing of the chimpanzee louse genome [40]. The availability of the human body and chimp louse genomes together with those of their hosts allowed Johnson et al. [40] to ask whether the parasites' DNA mutation rates and evolutionary divergence times differ from humans and chimpanzees. Comparisons were based on 1534 orthologous protein-coding genes that differed across all four genomes. Emerging results from this work indicated that DNA substitutions are, on average, 14 times faster in the body and head lice as compared to their mammalian hosts. Understanding the underlying

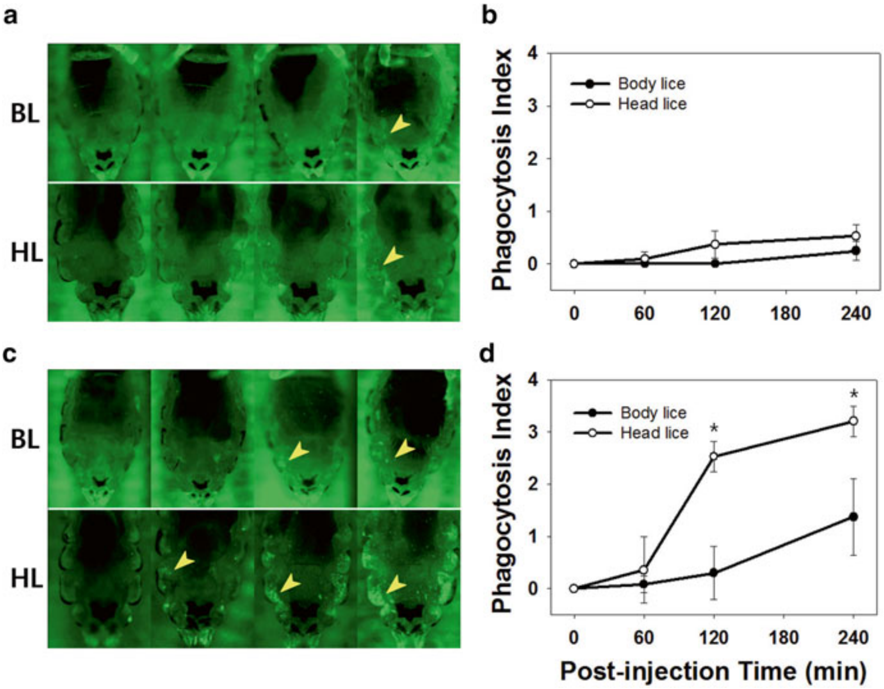


Fig. 1.4 Representative fluorescence microscopic images of abdominal region of the lice injected with FITC-labeled *Staphylococcus aureus* (**a**) or *Escherichia coli* (**c**) and the time course of phagocytosis as determined by phagocytosis index (**b** and **d**). The fluorescence mages were obtained from a single louse in a time series. *Arrowheads* in *panels***a** and **c** indicate typical phagocyte clusters immobilized in the lateral region of abdomen. The size of body louse images was reduced 1.5-fold to make it similar to that of head lice. In *panels***b** and **d**, symbols with *star mark* (*) indicate statistical difference between body and head lice ($p < 0.05$). *BL* body lice, *HL* head lice (Figure and figure legend are reproduced from Kim et al. [36] with permission from Elsevier)

mechanisms that drive these differences remains an interesting biological question for future studies.

1.7 Insecticide Toxicology

Annotation of the body louse genome revealed that detoxification genes (i.e., P450, GSTs, and EST) and ATP-binding cassette transporter (ABC transporter) genes are dramatically reduced in the body louse compared to other insects. The same set of orthologous genes was also found in the head louse by transcriptome analysis by Olds et al. [19]. The numbers of detoxification genes present in head and body lice were approximately half the number found in *D. melanogaster* and *Anopheles gambiae*. Despite the reduction in number, both head and body lice still retain at

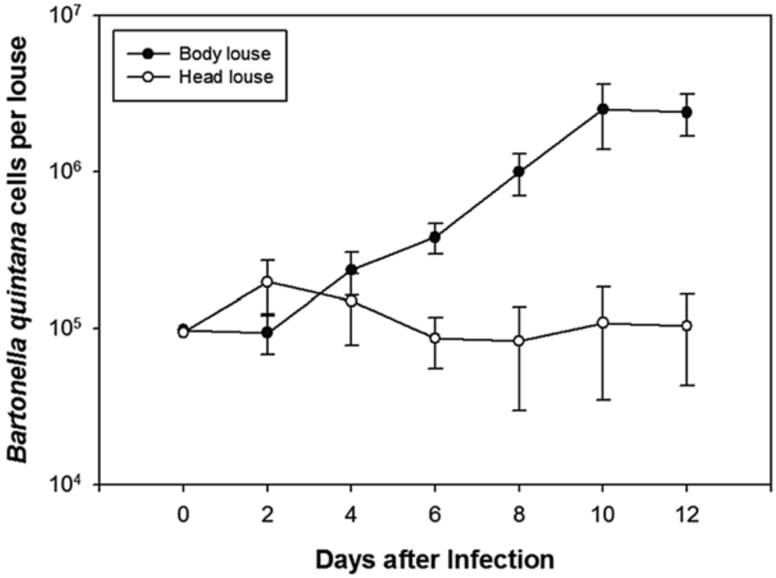


Fig. 1.5 *Bartonella quintana* proliferation in body and head lice. Adult female 1–4-day-old body (San Francisco strain, Frisco BL) and head (Bristol strain, BR-HL) lice maintained on the in vitro rearing system [38] were fed on human blood inoculated with *B. quintana* harvested from 7 to 10-day-old plates, at approximately 1×10^7 CFU/ml blood [39]. After an overnight feed, lice were transferred to uninfected blood. Five lice were collected at 2 day intervals and genomic DNA extracted to determine *B. quintana* proliferation, with primers amplifying an 89 base pair fragment in the 16S-23S rRNA gene of *B. quintana* [39]. Cell counts were normalized to the lowest 0 day cell count as well as relative lice gDNA detected by louse sodium channel primers (Figure and figure legend are reproduced from Previte et al. [37] with permission from the Royal Entomological Society)

least a minimum inventory of detoxification genes which can confer metabolic resistance to insecticides (e.g., clade 3 and 4 P450s, Delta GSTs, B clade Ests, and B/C subfamily ABC transporters), suggesting their high capability for resistance development. In fact, an Est gene (*HLCbE3*) was identified as a putative malathion carboxylesterase (MCE) through transcriptional profiling of five Est genes that were annotated to be catalytically active. Knockdown of *HLCbE3* transcription by RNA interference (RNAi) in a malathion-resistant strain increased malathion susceptibility, confirming its identity as an MCE responsible for malathion resistance in the head louse [41].

In an attempt to identify inducible metabolic factors that are involved in ivermectin tolerance and possible resistance in body lice, Yoon et al. [21] established a noninvasive induction assay (brief exposure to sublethal doses in a stress-reducing fashion that resulted in tolerance) and investigated the transcriptional profiles of P450 and ABC transporter genes after treatment with ivermectin. *CYP6CJ1*, *CYP9AG1*, *CYP9AG2*, and *PhABCC4* were determined to be most significantly overexpressed and to be most closely related to genes from other

organisms that metabolize insecticides, including ivermectin. Knockdown of *CYP9AG2* or *PhABCC4* by RNAi in noninduced female lice increased sensitivity to ivermectin, suggesting that these two genes are associated with its xenobiotic metabolism, thereby resulting in tolerance. Since genetic changes associated with some of these detoxification genes would most likely be involved in evolution of resistance following selection with pesticide, the noninvasive induction assay appears to be useful in screening inducible metabolic factors prior to resistance development and for employment in proactive resistance monitoring schemes.

1.8 Reverse Genetics: A Necessary Tool to Address Functional Genomic Questions

The use of genomic tools to understand transcriptional differences has been a major step forward in understanding the potential roles of specific genes in biological processes. By themselves, such studies only provide the basis for the development of hypotheses. In many cases, however, testing of these hypotheses needs to involve a selective knockdown in expression of a candidate gene, known as a reverse genetics approach. Thus, reverse genetics allows one to ask whether a specific gene is involved in a given biological process. RNAi takes advantage of the fact that eukaryotic cells are able to recognize double-stranded RNA (dsRNA) and then selectively degrade homologous RNA sequences, which in turn results in sequence-specific gene silencing [42]. For example, if a particular gene is thought to play a role in a given phenotype, e.g., a P450 conferring pesticide resistance, knockdown of its transcript through RNAi may result in the organism no longer (or temporarily) displaying that specific phenotype. For species with a large expansion of gene families where there might be numerous candidate genes, this can be an onerous task. A major advantage of using this approach in lice is that many gene families have restricted numbers of genes. Thus, a smaller set of genes (in a given family or class of genes) can be tested to determine which one is critical for a given biological process.

Two lines of evidence have been used to support the concept that RNAi can be used in lice as a reverse genetic tool. First, Pittendrigh et al. [18] showed bioinformatically, and then Yoon et al. [21] demonstrated, respectively, the existence of a complete set of RNAi machinery and an RNAi response in body lice (Table 1.2; [18]). Subsequently, Yoon et al. [21] obtained direct experimental evidence for the hypothesis that RNAi could be used to produce a measurable phenotypic change in body lice. Thus, lice in which *CYP9AG2* and *ABCC4* transcripts were knocked down by RNAi showed increased sensitivity to ivermectin (Fig. 1.6).

Although RNAi is currently being discussed in the scientific community as a strategy for insect control, many issues would have to be addressed before this could be considered for practical applications. First, how would dsRNA be delivered, and would that system result in effective insect control? Second, what are the implications and risks for nontargets (i.e., the human hosts), and what is the possibility of

Table 1.2 Genes identified in body louse that are homologous to members of the RNAi pathway in *Drosophila melanogaster*

Gene or microRNA name	Symbol	Flybase ID	Body louse	*E*-value
Argonaute 1	AGO1	FBgn002661	PHUM617260-PA	0.00E+00
Argonaute 2	AGO-2	FBgn0046812	PHUM004130-PA	1.00E−137
Argonaute 3	AGO-3	FBgn0250816	PHUM411830-PA	1.00E−162
Armitage	*armi*	FBgn0041164	PHUM074090-PA	1.00E−158
Ars2	Ars-2	FBgn0033062	PHUM507110-PA	0.00E+00
Aubergine	aub	FBgn0000146	PHUM563960-PA	0.00E+00
Capsuléen	csul	FBgn0015925	PHUM336830-PA	1.00E−125
Dicer-1	Dcr-1	FBgn0039016	PHUM435060-PA	0.00E+00
Dicer-2	Dcr-2	FBgn0034246	PHUM174480-PA	5.00E−78
Drosha	drosha	FBgn0026722	PHUM524860-PA	0.00E+00
Fmr1	Fmr1	FBgn0028734	PHUM440700-PA	1.00E−141
Gawky	gw	FBgn0051992	PHUM421980-PA	9.00E−76
Ge-1	Ge-1	FBgn0032340	PHUM249360-PA	3.00E−67
Hen1	Hen-1	FBgn0033686	PHUM430970-PA	1.00E−58
Loquacious	loqs	FBgn0032515	PHUM559590-PA	3.00E−90
Maternal expression at 31B	me31B	FBgn0004419	PHUM345900-PA	0.00E+00
microRNA encoding gene mir-14	mir-14	FBgn0046827	Present	N/A
Mir-184 (microRNA)	mir-184	FBgn0067726	Present	N/A
Partner of drosha	pasha	FBgn0039861	PHUM574170-PA	1.00E−160
piRNA methyltransferase	Hen-1/ pimmet	FBgn0033686	PHUM430970-PA	1.00E −58
Piwi	piwi	FBgn0004872	PHUM563960-PA	1.00E−167
r2d2	r2d2	FBgn0031951	PHUM504330-PA	1.00E−12
Twin	twin	FBgn0039168	PHUM129580-PA	5.00E−90
Zucchini	zuc	FBgn0004056	PHUM318600-PA	1.00E−16
Hsp90	Hsp90	FBgn0001233	PHUM581090-PA	0.00E+00
Rm62	Dmp68	FBgn0003261	PHUM521070-PA	1.00E−176
Vasa intronic gene	VIG	FBgn0024183	PHUM032630-PA	4.00E−24
Vasa intonic gene-2	VIG-2	FBgn0046214	PHUM032630-PA	9.00E−30
Spindle-E (homeless)	spn-E	FBgn0003483	PHUM492140-PA	0.00E+00
Yb	Yb	FBgn0000928	PHUM090360-PA	1.00E−16
Rhino	Rhino	FBgn0004400	PHUM169860-PA	3.00E−10
Elp1 (RNA-dependent RNApol)	Elp1	FBgn0037926	PHUM473060-PA	1.00E−136

Two microRNAs potentially associated with RNAi are also listed
From Pittendrigh et al. [18] with permission (Landes Biosciences)

successful approval of this approach by a regulatory agency? Finally, not all insect populations respond in the same way to RNAi treatment [43]. Thus, what is the potential for evolution of resistance to an RNAi-based control strategy in louse populations?

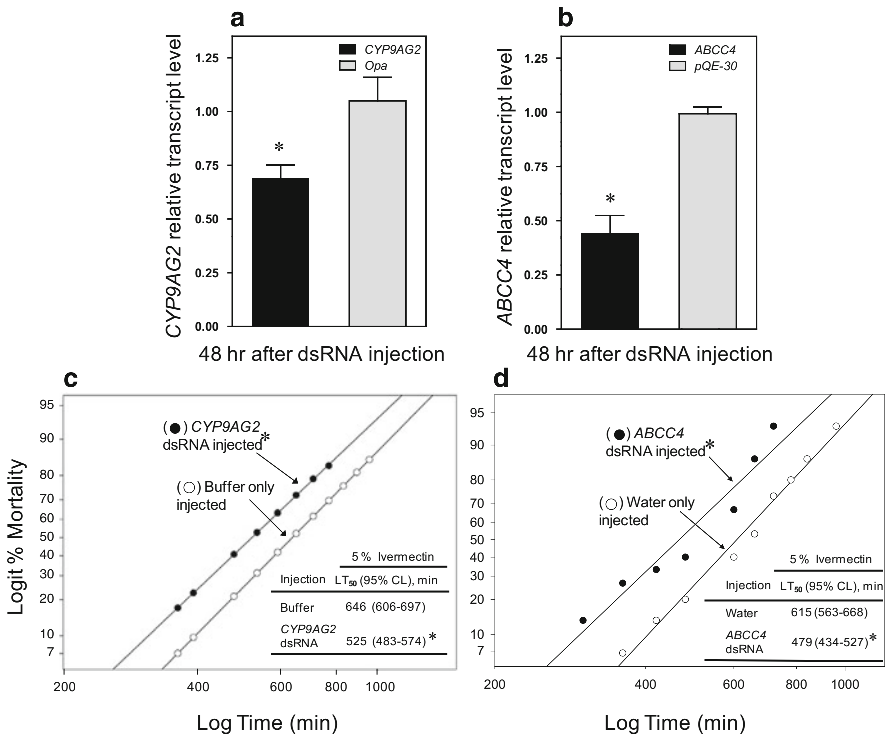


Fig. 1.6 Relative transcript levels (panels **a** and **b**) and mortality responses (panels **c** and **d**) of body louse females to a lethal contact amount of ivermectin (5 % IVM) following injection of dsRNA targeting either louse CYP9AG2 or ABCC4. Lice were also injected with either dsRNA of the odd-paired gene, opa (GeneBank accession no. S78339) for P450 silencing or with dsRNA of the *E. coli* plasmid, pQE30, for ABC transporter silencing as sham injected controls. *Asterisks* (*) in *panels* **a** and **b** indicate that *CYP9AG2* and *ABCC4* dsRNA significantly suppress the levels of *CYP9AG2* and ABCC4 transcripts, respectively (student's *t*-test, $P<0.05$). In *panel* **c**, the bioassay was started 48 h after *CYP9AG2* dsRNA injection. In *panel* **d**, the bioassay was started 12 h after ABCC4 dsRNA injection. *Asterisks* (*) in *panels* **c** and **d** indicate that the mortality responses of lice injected with dsRNAs were significantly different from their respective controls (buffer or water only injected, maximum log-likelihood ratio test, $P<0.05$) (Figure and legend are reproduced from Yoon et al. [21] with permission from the Royal Entomological Society)

Despite these potentially important applications in insect control and human health, due to the reduced number of genes in gene families associated with sensing and responding to the environment, the real potential of RNAi in lice involves using it as a tool to ask questions about the role of genes and gene products in specific biological processes.

1.9 Conclusion

Within less than a decade, the louse community has gone from having hardly any molecular data to a comprehensive genome project on body lice, an ORF analysis of head lice, extensive transcriptomic data providing testable hypotheses on important aspects of louse biology, and finally, an RNAi system to perform reverse genetic studies to test hypotheses about important health-related issues and fundamental insect biology and evolution that emerge from future transcriptomic and proteomic datasets. Thus, the stage has been set for the use of human body and head lice as system to address important biological questions that go well beyond just the interests of the louse research community.

References

1. Pedra JH, Brandt A, Li HM, Westerman R, Romero-Severson J, Pollack RJ, Murdock LL, Pittendrigh BR (2003) Transcriptome identification of putative genes involved in protein catabolism and innate immune response in human body louse (Pediculicidae: *Pediculus humanus*). Insect Biochem Mol Biol 33:1135–1143
2. Durden LA, Musser GG (1994) The sucking lice (Insecta: Anoplura) of the world: a taxonomic checklist with records of mammalian hosts and geographic distributions. B Am Mus Nat Hist 218:1–90
3. Khudobin VV (1995) The adaptive potentials of human head and clothes lice when parasitizing on man. Med Parazitol (Mosk) 1:23–25
4. Busvine JR (1948) The head and body races of *Pediculus humanus* L. Parasitology 39:1–16
5. Schaefer CW (1978) Ecological separation of the human head lice and body lice (Anoplura: Pediculidae). Trans R Soc Trop Med Hyg 72:669–670
6. Bacot A (1917) A contribution to the bionomics of *Pediculus humanus* (Vestimenti) and *Pediculus capitis*. Parasitology 9:228–258
7. Mullen GR, Durden LA (2009) Medical and veterinary entomology, 2nd edn. Academic/Elsevier Science, San Diego
8. Light JE, Toups MA, Reed DL (2008) What's in a name: the taxonomic status of human head and body lice. Mol Phylogenet Evol 47:1203–1216
9. Bonilla DL, Kabeya H, Henn J, Kramer VL, Kosoy MY (2009) *Bartonella quintana* in body lice and head lice from homeless persons, San Francisco, California, USA. Emerg Infect Dis 15:912–915
10. Toloza AC, Vassena C, Picollo MI (2008) Ovicidal and adulticidal effects of monoterpenoids against permethrin-resistant human head lice, *Pediculus humanus capitis*. Med Vet Entomol 22:335–339
11. Sasaki T, Poudel SK, Isawa H, Hayashi T, Seki N, Tomita T, Sawabe K, Kobayashi M (2006) First molecular evidence of *Bartonella quintana* in *Pediculus humanus capitis* (Phthiraptera: Pediculidae), collected from Nepalese children. J Med Entomol 43:110–112
12. Rhee KY, Johnson Jr WD (2009) *Borrelia* species (relapsing fever). In: Mandell GL, Bennett JE, Dolan R (eds) Mandell, Douglas, and Bennett's principles and practice of infectious disease. Lancet infectious diseases. Churchill Livingstone/Elsevier, Philadelphia, pp 3067–3070
13. Johnston JS, Yoon KS, Strycharz JP, Pittendrigh BR, Clark JM (2007) Body lice and head lice (Anoplura: Pediculidae) have the smallest genomes of any hemimetabolous insect reported to date. J Med Entomol 44:1009–1012

14. Pittendrigh BR, Clark JM, Johnston JS, Lee SH, Romero-Severson J, Dasch GA (2005) Proposal for the sequencing of a new target genome: white paper for a human body louse genome project. National Human Genome Research Institute. www.genome.gov/10002154
15. Pittendrigh BR, Clark JM, Johnston JS, Lee SH, Romero-Severson J, Dasch GA (2006) Sequencing of a new target genome: the *Pediculus humanus humanus* (Phthiraptera: Pediculidae) genome project. J Med Entomol 43:1103–1111
16. Kirkness EF, Haas BJ, Sun W, Braig HR, Perotti MA, Clark JM, Lee SH, Robertson HM, Kennedy RC, Elhaik E, Gerlach D, Kriventseva EV, Elsik CG, Graur D, Hill CA, Veenstra JA, Walenz B, Tubio JM, Ribeiro JM, Rozas J, Johnston JS, Reese JT, Popadic A, Tojo M, Raoult D, Reed DL, Tomoyasu Y, Kraus E, Mittapalli O, Margam VM, Li HM, Meyer JM, Johnson RM, Romero-Severson J, Vanzee JP, Alvarez-Ponce D, Vieira FG, Aguade M, Guirao-Rico S, Anzola JM, Yoon KS, Strycharz JP, Unger MF, Christley S, Lobo NF, Seufferheld MJ, Wang N, Dasch GA, Struchiner CJ, Madey G, Hannick LI, Bidwell S, Joardar V, Caler E, Shao R, Barker SC, Cameron S, Bruggner RV, Regier A, Johnson J, Viswanathan L, Utterback TR, Sutton GG, Lawson D, Waterhouse RM, Venter JC, Strausberg RL, Berenbaum MR, Collins FH, Zdobnov EM, Pittendrigh BR (2010) Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. Proc Natl Acad Sci U S A 107:12168–12173
17. Lee SH, Kang JS, Min JS, Yoon KS, Strycharz JP, Johnson R, Mittapalli O, Margam VM, Sun W, Li HM, Xie J, Wu J, Kirkness EF, Berenbaum MR, Pittendrigh BR, Clark JM (2010) Decreased detoxification genes and genome size make the human body louse an efficient model to study xenobiotic metabolism. Insect Mol Biol 19:599–615
18. Pittendrigh BR, Berenbaum MR, Seufferheld MJ, Margam VM, Strycharz JP, Yoon KS, Sun W, Reenan R, Lee SH, Clark JM (2011) Simplify, simplify: lifestyle and compact genome of the body louse provide a unique functional genomics opportunity. Commun Integr Biol 4:188–191
19. Olds BP, Coates BS, Steele LD, Sun W, Agunbiade TA, Yoon KS, Strycharz JP, Lee SH, Paige KN, Clark JM, Pittendrigh BR (2012) Comparison of the transcriptional profiles of head and body lice. Insect Mol Biol 21:257–268
20. Drali R, Boutellis A, Raoult D, Rolain JM, Brouqui P (2013) Distinguishing body lice from head lice by multiplex real-time PCR analysis of the Phum_PHUM540560 gene. PLoS One 8:e58088
21. Yoon KS, Strycharz JP, Baek JH, Sun W, Kim JH, Kang JS, Pittendrigh BR, Lee SH, Clark JM (2011) Brief exposures of human body lice to sublethal amounts of ivermectin over-transcribes detoxification genes involved in tolerance. Insect Mol Biol 20:687–699
22. Gordon KH, Waterhouse PM (2007) RNAi for insect-proof plants. Nat Biotechnol 25:1231–1232
23. Puchta O (1955) Experimental studies on the significance of symbiosis in the clothes louse *Pediculus vestimenti* Burm. Z Parasitenkd 17:1–40
24. Bai X, Mamidala P, Rajarapu SP, Jones SC, Mittapalli O (2011) Transcriptomics of the bed bug (*Cimex lectularius*). PLoS One 6:e16336
25. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, Dijkstra MB, Oettler J, Comtesse F, Shih CJ, Wu WJ, Yang CC, Thomas J, Beaudoing E, Pradervand S, Flegel V, Cook ED, Fabbretti R, Stockinger H, Long L, Farmerie WG, Oakey J, Boomsma JJ, Pamilo P, Yi SV, Heinze J, Goodisman MA, Farinelli L, Harshman K, Hulo N, Cerutti L, Xenarios I, Shoemaker D, Keller L (2011) The genome of the fire ant *Solenopsis invicta*. Proc Natl Acad Sci U S A 108:5679–5684
26. Wei DD, Chen EH, Ding TB, Chen SC, Dou W, Wang JJ (2013) De novo assembly, gene annotation, and marker discovery in stored-product pest *Liposcelis entomophila* (Enderlein) using transcriptome sequences. PLoS One 8:e80046
27. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, Denas O, Elhaik E, Fave MJ, Gadau J, Gibson JD, Graur D, Grubbs KJ, Hagen DE, Harkins TT, Helmkampf M, Hu H, Johnson BR, Kim J, Marsh SE, Moeller JA,

Munoz-Torres MC, Murphy MC, Naughton MC, Nigam S, Overson R, Rajakumar R, Reese JT, Scott JJ, Smith CR, Tao S, Tsutsui ND, Viljakainen L, Wissler L, Yandell MD, Zimmer F, Taylor J, Slater SC, Clifton SW, Warren WC, Elsik CG, Smith CD, Weinstock GM, Gerardo NM, Currie CR (2011) The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle. PLoS Genet 7:e1002007

28. You M, Yue Z, He W, Yang X, Yang G, Xie M, Zhan D, Baxter SW, Vasseur L, Gurr GM, Douglas CJ, Bai J, Wang P, Cui K, Huang S, Li X, Zhou Q, Wu Z, Chen Q, Liu C, Wang B, Li X, Xu X, Lu C, Hu M, Davey JW, Smith SM, Chen M, Xia X, Tang W, Ke F, Zheng D, Hu Y, Song F, You Y, Ma X, Peng L, Zheng Y, Liang Y, Chen Y, Yu L, Zhang Y, Liu Y, Li G, Fang L, Li J, Zhou X, Luo Y, Gou C, Wang J, Wang J, Yang H, Wang J (2013) A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet 45:220–225
29. Hansson BS, Stensmyr MC (2011) Evolution of insect olfaction. Neuron 72:698–711
30. Pakpour N, Riehle MA, Luckhart S (2014) Effects of ingested vertebrate-derived factors on insect immune responses. Curr Opin Insect Sci 3:1–5
31. Glastad KM, Hunt BG, Yi SV, Goodisman MAD (2011) DNA methylation in insects: on the brink of the epigenomic era. Insect Mol Biol 20:553–565
32. Sztal T, Chung H, Berger S, Currie PD, Batterham P, Daborn PJ (2012) A cytochrome p450 conserved in insects is involved in cuticle formation. PLoS One 7:e36544
33. Russell RJ, Scott C, Jackson CJ, Pandey R, Pandey G, Taylor MC, Coppin CW, Liu J-W, Oakeshott JG (2011) The evolution of new enzyme function: lessons from xenobiotic metabolizing bacteria versus insecticide-resistant insects. Evol Appl 4:225–248
34. Zhu JY, Zhao N, Yang B (2012) Global transcriptional analysis of olfactory genes in the head of pine shoot beetle, *Tomicus yunnanensis*. Comp Funct Genomics 2012:491748
35. Vieira FG, Rozas J (2011) Comparative genomics of the odorant-binding and chemosensory protein gene families across the Arthropoda: origin and evolutionary history of the chemosensory system. Genome Biol Evol 3:476–490
36. Kim JH, Min JS, Kang JS, Kwon DH, Yoon KS, Strycharz J, Koh YH, Pittendrigh BR, Clark JM, Lee SH (2011) Comparison of the humoral and cellular immune responses between body and head lice following bacterial challenge. Insect Biochem Mol Biol 41:332–339
37. Previte D, Olds BP, Yoon K, Sun W, Muir W, Paige KN, Lee SH, Clark J, Koehler JE, Pittendrigh BR (2014) Differential gene expression in laboratory strains of human head and body lice when challenged with *Bartonella quintana*, a pathogenic bacterium. Insect Mol Biol 23:244–254
38. Yoon KS, Strycharz JP, Gao J-R, Takano-Lee M, Edman JD, Clark JM (2006) An improved *in vitro* rearing system for the human head louse allows the determination of resistance to formulated pediculicides. Pestic Biochem Physiol 86:195–202
39. Seki N, Kasai S, Saito N, Komagata O, Mihara M, Sasaki T, Tomita T, Sasaki T, Kobayashi M (2007) Quantitative analysis of proliferation and excretion of *Bartonella quintana* in body lice, *Pediculus humanus* L. Am J Trop Med Hyg 77:562–566
40. Johnson KP, Allen JM, Olds BP, Mugisha L, Reed DL, Paige KN, Pittendrigh BR (2014) Rates of genomic divergence in humans, chimpanzees and their lice. Proc Biol Sci 281:20132174
41. Kwon DH, Kim JH, Kim YH, Yoon KS, Clark JM, Lee SH (2014) Identification and characterization of an esterase involved in malathion resistance in the head louse *Pediculus humanus capitis*. Pestic Biochem Physiol 112:13–18
42. Hannon GJ (2002) RNA interference. Nature 418:244–251
43. Chu CC, Sun W, Spencer JL, Pittendrigh BR, Seufferheld MJ (2014) Differential effects of RNAi treatments on field populations of the western corn rootworm. Pestic Biochem Physiol 110:1–6

Chapter 2
Advances in the Genomics of the Whitefly *Bemisia tabaci*: An Insect Pest and a Virus Vector

Surapathrudu Kanakala and Murad Ghanim

Abstract The sweetpotato whitefly, *Bemisia tabaci*, is a devastating cosmopolitan insect pest that inflicts serious damage by direct feeding on plants, secreting honeydew, and vectoring more than 100 plant viruses that belong to different virus genera. The interactions between the whitefly and plant viruses, plants, and environmental factors have been extensively studied. In recent years more than 100,000 expressed sequences tags (ESTs) from the whitefly have been made available to the scientific community by several mass sequencing projects, and a genome sequencing project is underway. Tools for functional analysis of gene expression are being developed for studies in the whitefly. Combining EST and genomic sequences with functional analysis will pave the way for addressing urgent issues in whitefly research and developing better strategies for whitefly control.

Abbreviations

AbMV	Abutilon mosaic virus
CabLCV	Cabbage leaf curl virus
CP	Coat protein
EST	Expressed sequence tag
FISH	Fluorescence in situ hybridization
HSP	Heat shock protein
ITS1	Ribosomal internal transcribed spacer 1
JA	Jasmonic acid
MEAM1	Middle East-Asia Minor 1
MED	Mediterranean
mtCO1	Mitochondrial cytochrome oxidase 1
qRT-PCR	Quantitative real-time RT-PCR

S. Kanakala • M. Ghanim (✉)
Department of Entomology, Volcani Center, 6, Bet Dagan 50250, Israel
e-mail: kanakalavit@gmail.com; ghanim@agri.gov.il

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*,
Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_2

RAPD-PCR	Random amplified polymorphic cDNA-polymerase chain reaction
RFLP	Restriction fragment length polymorphism
SA	Salicylic acid
SCAR	Sequence-characterized amplified regions
ToMoV	Tomato mottle virus
TYLCCNB	Tomato yellow leaf curl China virus betasatellite
TYLCCNV	Tomato yellow leaf curl China virus
TYLCSV	Tomato yellow leaf curl Sardinia virus
TYLCV	Tomato yellow leaf curl virus
WmCSV	Watermelon chlorotic stunt virus

2.1 Introduction

For decades, agriculture has been the main source of food and other useful products for humans and livestock. The world population is expected to grow by one-third from the present by 2050 (www.fao.org) leading to an increase in demand for agricultural products to feed growing populations worldwide. Production of agricultural crops is at risk due to the incidence of herbivores and disease-causing pathogens. Herbivores, mainly insect pests, are undoubtedly the most important competitors for food, fiber, and other natural products. Beyond their economic importance, they have a direct impact on agricultural food production by chewing, sucking, and boring plant parts and, most importantly, by spreading devastating plant pathogens, particularly plant viruses. Many important plant viruses depend on insect vectors which mainly belong to the orders *Hemiptera*, *Thysanoptera*, and *Coleoptera*, which transmit a great diversity of plant viruses across the tropical and subtropical regions of the world. Among them, phloem-feeding insects (aphids, whiteflies, plant, and leafhoppers) are the most common insect vectors that transmit numerous viruses and cause significant losses to food crops of socioeconomic importance.

In the last 20 years of the twentieth century, vector biology research switched from an understanding of the general biology of virus-vector relationships to a more deep understanding of the viral and vector proteins and components involved in the transmission process. However, the insect vector component was largely ignored due to a lack of genetic and genomic resources. Developments in mass sequencing projects such as expressed sequence tags (ESTs), together with genome annotation and functional gene discovery, are greatly advancing our understanding of the transmission process of viruses by their vectors and the components required to facilitate virus transmission. The emergence of massive sequencing technologies such as 454 pyrosequencing and the Illumina sequencing platform has been widely used to understand the underlying molecular mechanisms [1] involved in virus transmission, mating behavior, and genomes of many hemipterans and the whitefly, *Bemisia tabaci*, its endosymbionts, and the interactions of the insect with the environment. Functional genomic projects of the whitefly will facilitate an understanding of not only its interactions with viruses but also its resistance to insecticides, its develop-

mental patterns, and its interactions with plants and other microorganisms. This chapter reviews the recent advances in the genomics of *B. tabaci* and the relevance of these advancements to our understanding of its interactions with various factors.

2.2 Status of *B. tabaci* as a Global Agricultural Pest

The whitefly, *B. tabaci* (Gennadius) (*Hemiptera*: *Aleyrodidae*), is a phloem-feeding insect that inflicts a serious worldwide threat by direct feeding on plants and by the transmission of plant viruses [2–4] (Fig. 2.1). *B. tabaci* was first described in 1889 as a tobacco pest in Greece and named *Aleyrodes tabaci* [5]; later on it was described as *A. inconspicua* Quaintance in 1900 [6] and *B. tabaci* in 1957 [7]. Comprehensive recent reviews about the history of *B. tabaci* [8] and its cryptic species status [9–13] are available.

B. tabaci adults are about 1–2 mm long and have opaque-shaped wings covered with a whitish powder or wax, leading to the common name of whiteflies. *B. tabaci* was first reported to be a serious insect pest in the late 1920s in Northern India [14, 15] and is now globally distributed, in the USA, Africa, the Middle East, Europe,

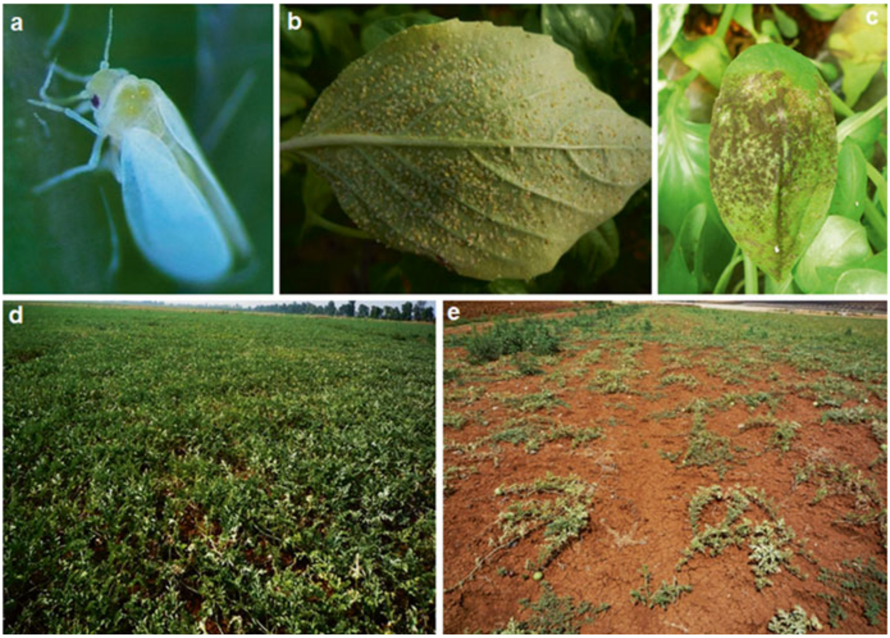


Fig. 2.1 Direct and indirect plant damage caused by *Bemisia tabaci*. (**a**) *B. tabaci* adult. (**b**) Representative damage caused by *B. tabaci* feeding showing an infested basil leaf with *B. tabaci* nymphs. (**c**) A basil leaf showing sooty mold growth after *B. tabaci* infestation and massive secretion of honeydew. (**d**) An uninfested watermelon field. (**e**) Another watermelon field that was infested by two *B. tabaci*-transmitted begomoviruses *squash leaf curl virus* (SLCV) and *watermelon chlorotic stunt virus* (WmCSV)

the Orient, Russia, Southeast Asia, and South America [3], except in Antarctica [16]. It is a highly polyphagous insect that feeds on over 700 plant species from 86 families, including a large number of agriculturally and industrially important crops [9, 17, 18]. Whiteflies damage plants in various ways. The delicate mouthparts of the whitefly adults and nymphal stages cause direct feeding damage by weakening plants (Fig. 2.1). They also cause indirect damage due to the secretion of honeydew that causes sooty mold growth which interferes with photosynthesis and makes the fruits become unsightly.

The whitefly's major impact on agriculture is due to the fact that this insect serves as a vector of more than 110 plant viruses [9, 16–19], many of which have great impact on plants and agricultural production. Estimates of their destructive reach extend to 20 million hectares of crops and 15 million farmers [20]. Next to aphids, whiteflies are of greater importance as vectors of the most devastating plant virus groups, including *Begomovirus*, *Crinivirus*, *Closterovirus*, *Carlavirus*, and *Torradovirus* [17, 21]. Among whitefly-transmitted viruses, 90 % belong to the *Begomovirus* genera [17], which include approximately 200 species, and have emerged as the most threatening group of plant viruses globally during the past two decades as reported from dicotyledonous host-causing diseases of economic importance.

B. tabaci is considered as a cryptic species complex [11]. Individual cryptic species within the complex differ in their adaptability to hosts, in their resistance to chemicals, and importantly in their ability to transmit begomoviruses [22–26]. Exotic whitefly biotypes or cryptic species transmit Old and New World begomoviruses in many places of the world resulting from changes in environmental conditions and agricultural practices. Studies on whitefly biotypes/cryptic species and the begomoviruses they transmit worldwide reveal that grouping of begomoviruses based on CP (coat protein) sequences parallels grouping of biotypes/genotypes so that both begomoviruses and vector are grouped in similar patterns according to their geographic origin. As different members of the species complex are morphologically indistinguishable, various molecular methods have been applied over the past two decades to delimit the members of the *B. tabaci* species complex.

2.3 Biotypes or Cryptic Species of *B. tabaci*: Definitions and Distribution

B. tabaci has numerous biotypes which transmit a large number of viruses that infect many important agricultural plants, causing major economic impacts [27]. As early as the 1950s, biotypes were proposed to characterize the morphologically indistinguishable populations of *B. tabaci* on the basis of host range, host plant adaptability, and plant virus transmission capabilities [28–30]. During the 1980s, the A biotype became a serious problem in cotton and cucurbits in the Southwestern USA and Mexico [3]. In 1991, the introduced B biotype displaced the A biotype in the Southwestern USA [3]. In the same year, Perring et al. [31] suggested that A and

B biotypes were separate species. A newly evolved B biotype, commonly referred to as the silver leaf whitefly or poinsettia strain, has a very wide host range, which has contributed to the spread of geminiviruses to new hosts [32] and has been documented as a separate species named *B. argentifolii* [33]. Using allozymes and the random amplified polymorphic cDNA-polymerase chain reaction (RAPD-PCR), Perring et al. [34] and Gawel and Bartlett [35] further showed consistent differences between A and B biotypes.

As the different members of the *B. tabaci* species complex are morphologically indistinguishable, various molecular methods have been applied over the past two decades to delimit the members of this species complex. The most popular techniques and the types of DNA markers used to study *B. tabaci* are sequence characterized amplified regions (SCAR) [36–41], cleaved amplified polymorphic sequences or restriction fragment length polymorphisms (CAPS/RFLP) [37, 42], amplified fragment length polymorphisms (AFLP) [43], 16S ribosomal RNA [44], mitochondrial cytochrome oxidase 1 (mtCO1) [10–12, 45–49], nuclear ribosomal internal transcribe spacer 1 (ITS1) [50–54], and microsatellites [55–65]. Among them, mtCOI has several advantages over other approaches and has been extensively used [11].

Using CO1-based Bayesian phylogenetic analysis and sequence divergence, Dinsdale et al. [10] and De Barro et al. [11] proposed a speciation system on the basis of a demarcation criterion of a 3.5 % divergence threshold. Recently, Lee et al. [49] observed that a 4.0 % genetic boundary was more realistic than 3.5 % in distinguishing *B. tabaci* species. Following the above criteria, 37 morphologically indistinguishable species (Africa, Asia I, Asia II 1, Asia II 2, Asia II 3, Asia II 4, Asia II 5, Asia II 6, Asia II 7, Asia II 8, Asia II 9, Asia II 10, Asia II 11, Asia II 12, Asia III, Asia IV, Australia, Australia/Indonesia, China 1, China 2, China 3, China 4, Indian Ocean, Middle East-Asia Minor (MEAM) I, MEAM II, Mediterranean (MED), New World 1, New World 2, Japan 1, Japan 2, Uganda, Italy 1, Sub-Saharan Africa 1, Sub-Saharan Africa 2, Sub-Saharan Africa 3, Sub-Saharan Africa 4, Sub-Saharan Africa 5) have been currently delimited at the global level [10–13, 46–49, 66]. Among the *B. tabaci* species complex, the most important biotypes worldwide are B and Q, recently termed as the MEAM1 and MED species, respectively [3, 11, 44, 67].

2.4 Genomic and Postgenomic Resources Developed for *B. tabaci* and What Is Missing

As indicated, *B. tabaci* is a complex of biotypes that differ in their behavior, plant host, ability to induce disorders in plants, ability to transmit plant viruses, the bacterial endosymbionts they harbor, and genetic makeup [68]. In addition to the endogenous species, new invasive and better fit biotypes such as B and Q have invaded crop systems, exacerbating damage [69]. Very little is known about the virus transmission specificity, mating behavior, or genomes of either the whitefly or its bacterial endosymbionts. Collective functional genome projects of the whitefly will help to understand the whitefly genetic makeup and its interactions with environmental factors.

In 2005, the first attempt to measure nuclear DNA content of *B. tabaci* complex males and females was estimated using flow cytometry. This was the first step toward the exploration of the whitefly genome [70], and yielded values of 1.04 and 2.06 pg, respectively. Conversion between DNA content and genome size (1 pg DNA = 980 Mbp) indicated that the haploid genome size of *B. tabaci* is 1020 Mbp, which is approximately five times the size of the genome of the fruit fly, *Drosophila melanogaster* [70].

The first large-scale sequencing of ESTs from *B. tabaci* was initiated by Leshkowitz et al. in 2006 [71]. To address whitefly genomic sequence information, they constructed three cDNA libraries from non-viruliferous whiteflies (eggs, immature instars, and adults) and two libraries from adult insects with *tomato yellow leaf curl virus* (TYLCV) and *tomato mottle virus* (ToMoV). In total, 9110 sequences from the libraries were found to be involved in cellular and developmental processes. In addition, approximately 1000 bases were aligned with the genome of the *B. tabaci* primary endosymbiotic bacterium, *Candidatus* Portiera aleyrodidarum, originating primarily from the egg and immature instar libraries [71]. Apart from the mitochondrial sequences, abundant sequences in the libraries encoded vitellogenin gene sequences, indicating that much of the gene expression in *B. tabaci* is directed toward the production of eggs [71]. Studies thereafter were aimed to understand the mechanisms and the function of genes involved in various aspects of the whitefly interaction with its environment, and the first case was its resistance to insecticides. For the first time a DNA microarray containing 6000 unique ESTs from the whitefly was developed in 2007, and deep analysis revealed ESTs involved in insecticide resistance and xenobiotic detoxification such as those transcribed from P450 monooxygenases and oxidative stress genes, genes associated with protein, lipid and carbohydrate metabolism, and others related to juvenile hormone-associated processes in insects such as oocyte and egg development [72].

Following these pioneering studies and the development of new deep sequencing technologies and bioinformatic tools, the first de novo characterization of a whitefly transcriptome without prior genome annotation was sequenced by Wang et al. in 2010 [73], using the Illumina sequencing platform. This study obtained 168,900 unique sequences and a large number of genes associated with specific developmental stages and insecticide resistance. In 2011, using the same technology, Luan et al. [74] investigated the transcriptional response of the invasive *B. tabaci* MEAM1 species to *tomato yellow leaf curl China virus* (TYLCCNV). Results showed that 1606 genes involved in 157 biochemical pathways were differentially expressed in the viruliferous whiteflies. This indicates that TYLCCNV can perturb the cell cycle and primary metabolism, which explains the negative effect of this virus on the longevity and fecundity of *B. tabaci* [75]. Further results showed that TYLCCNV can activate whitefly immune responses, such as autophagy and antimicrobial peptide production, which might lead to a gradual decrease of viral particles within the body of the infected insect. Furthermore, PCR results showed that TYLCCNV can invade ovary and fat body, and LysoTracker and Western blot analyses revealed that the invasion of TYLCCNV induced autophagy in both types of tissue. Surprisingly, TYLCCNV also suppressed whitefly immune responses by downregulating the

expression of genes involved in Toll-like signaling and mitogen-activated protein kinase (MAPK) pathways [74].

Using Illumina sequencing, further research investigated the transcriptome of the primary salivary glands of the MED species of the *B. tabaci* complex [76]. This study obtained 13,615 unigenes involved in metabolism and transport. Further analysis revealed genes related to processing of secretory proteins, secretion, and virus transmission [76]. Following this sequencing, Wang et al. [77, 78] sequenced an indigenous species, Asia II 3, and compared its genetic divergence with the transcriptomes of two invasive whitefly species, MEAM1 and MED. This study revealed a conserved group of 3203 protein families among the Asia II 3, MEAM1, and MED species which might be responsible for core cellular and physiological functions of the *B. tabaci* complex. Further results identifying hundreds of highly diverged genes and compiling sequence annotation data into functional groups found the most divergent gene classes to be the cytochrome P450 monooxygenases, glutathione metabolism, and oxidative phosphorylation. Moreover, many of these genes are predicted to be involved in protein metabolism such as peptide deformylase, cathepsin, cysteine proteinase, and metalloendopeptidase, which might be the driving force of the MEAM1-MED divergence and MEAM1- Asia II 3 differentiation [78]. A recent pyrosequencing study of the transcriptome of an invasive B biotype from China revealed a highly diverse bacterial community and robust system for insecticide resistance [79]. As mentioned above, de novo assembly generated 178,669 unigenes including 30,980 from the insects and 17,881 from the bacteria. Further in-depth transcriptome analysis revealed additional genes potentially involved in insecticide resistance and nutrient digestion [79].

Recently, Wang et al. [80] obtained the complete mitochondrial genome of the invasive Mediterranean species and identified 37 genes, including 13 protein-coding genes (PCGs), 2 ribosomal RNAs, and 22 transfer RNAs (tRNA). Comparative analyses of the genomes from MED and New World species revealed that there are no gene arrangements. This study also revealed that *atp6* and *atp8*, *nd4* and *nd4l*, and *nd6* and *cytb* were on the same cistronic transcripts, whereas the other mature mitochondrial transcripts were monocistronic.

Recent transcriptome profiling of *B. tabaci* revealed stage-specific gene expression signatures for thiamethoxam resistance [81, 82]. Furthermore, analysis of gut transcriptomes of two invasive whitefly species in the *B. tabaci* complex, MEAM1 and MED, demonstrated the important role of the gut in the metabolism of insecticides and secondary plant chemicals [83]. Interestingly, using a microarray from *B. tabaci* that contained about 6000 genes, it was shown that parasitization by the wasp, *Eretmocerus mundus*, induced genes related to immune responses and revealed a role for the bacterial symbiont, *Rickettsia*, in these responses [84].

The foregoing transcriptomic studies reveal that a number of highly expressed genes that belong to different and conserved metabolic pathways with other insects are involved in virus transmission, insecticide resistance, and immune responses to parasitoids. Taken together, the transcriptomic results collected so far provide not only a roadmap for further functional genomic studies and extensive whitefly research in general but also a large collection of gene and EST sequences for future genome sequencing and annotation efforts.

2.5 *B. tabaci* Bacterial Symbionts

Bacterial symbionts in insects have drawn the attention of many research projects around the world because of their potential for developing friendly pest control methods by intervention with the insect-symbiont interactions. Recent surveys of facultative endosymbionts in whiteflies and aphids have used hundreds of insect populations collected around the world [85–90]. *B. tabaci* exhibits the highest diversity of secondary endosymbionts that infect one insect, and populations were reported to harbor seven different facultative (secondary) symbionts including the *Alphaproteobacteria Rickettsia* (*Rickettsiales*), *Orentia*-like organism (*Rickettsiales*), *Wolbachia* (*Rickettsiales*), the *Gammaproteobacteria Arsenophonus* (*Enterobacteriales*), *Hamiltonella* (*Enterobacteriales*), *Cardinium* (*Bacteroidetes*), and *Fritschea* (*Chlamydiales*) [91–97] (Fig. 2.2). All whitefly species including *B. tabaci* biotypes harbor the primary symbiont *Portiera*; however, they differ in their secondary symbiont composition [91, 98]. A study from Israel showed that B and Q biotypes differ in their infection with secondary symbionts, and while the B biotype was infected with *Hamiltonella*, the Q biotype was infected with *Arsenophonus* and *Wolbachia.* Both biotypes were infected with *Rickettsia* and none was infected with *Fritschea* and *Cardinium* [85]. A different study surveyed the distribution of secondary symbionts in whitefly species collected in the Balkan region and detected only the Q biotype of *B. tabaci*, which harbored *Hamiltonella*, *Rickettsia*, *Wolbachia*, and *Cardinium*, while *Arsenophonus* and *Fritschea* were not detected in any *B. tabaci* populations [86]. Recent studies focused on the role that these symbionts might play in the biology of their whitefly hosts. Two studies showed that *Wolbachia* [99] and *Hamiltonella* [100, 101] increase the fitness and enhance *B. tabaci* MED biotype performance. Other studies have shown that *Rickettsia* increases *B. tabaci* tolerance to heat [102] and influences the whitefly response to insecticides [103, 104]. A study conducted in the USA showed that *Rickettsia* increases the fitness of the B biotype of *B. tabaci* and enhances the proportion of females in the populations [105].

Bacterial symbionts in *B. tabaci* have been shown to play an important role in begomovirus transmission [106, 107]. Gottlieb et al. [108] showed that only a GroEL protein from *Hamiltonella* interacted with TYLCV CP and protected the virus while circulating in the whitefly hemolymph for transmission, whereas other GroEL proteins from other symbionts did not interact with TYLCV CP. It was further shown that release of virions protected by GroEL occurs adjacent to the primary salivary gland. However, CP of *abutilon mosaic virus* (AbMV), a begomovirus non-transmissible by *B. tabaci*, was also found to bind to GroEL. The midgut was shown to serve as the barrier for AbMV transmission and not interactions with GroEL [107]. A study from India showed that GroEL protein of *Arsenophonus* interacted with the *cotton leaf curl virus* (CLCuV) CP, suggesting the involvement of *Arsenophonus* in the transmission of CLCuV in the Asia II genetic group of *B. tabaci* [109]. Additionally, a recent study has demonstrated that *Rickettsia* plays an important role in TYLCV transmission. *Rickettsia*-infected females transmitted TYLCV at twice the rate of *Rickettsia*-uninfected females and retained the virus longer [110]. Interestingly, an antagonistic relationship was discovered between

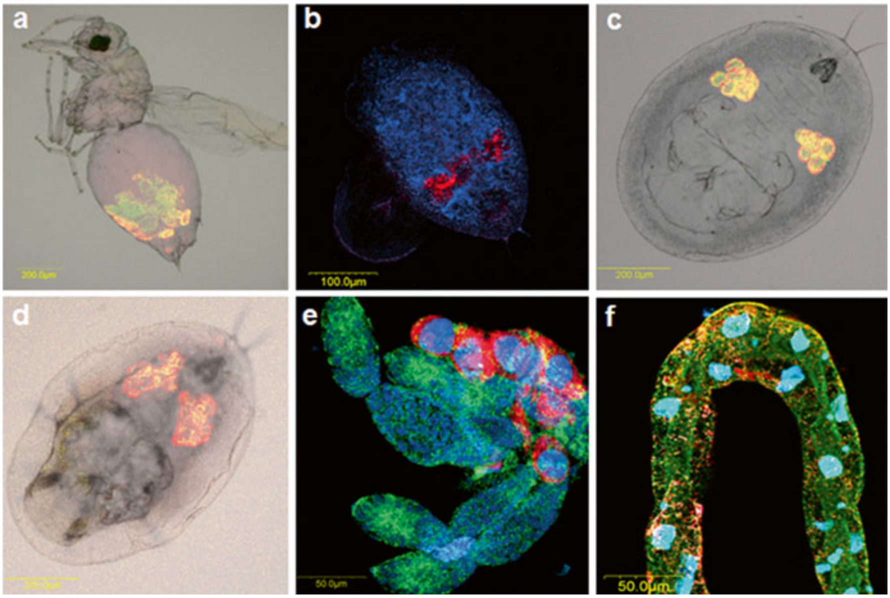


Fig. 2.2 Fluorescence in situ hybridization (FISH) analysis on representative *B. tabaci* developmental stages and dissected organs for the localization of representative symbiotic bacteria. (**a**) An adult whitefly showing the primary symbiont *Portiera* in *red* and the secondary symbiont *Hamiltonella* in *green*. (**b**) A second-stage nymph showing the primary symbiont *Portiera* in *red* and the secondary symbiont *Rickettsia* in *blue*. (**c**) A fourth-stage nymph showing the primary symbiont *Portiera* in *red* and the secondary symbiont *Hamiltonella* in *green*. (**d**) A pupa showing the primary symbiont *Portiera* in *red* and the secondary symbiont *Arsenophonus* in *yellow*. (**e**) A dissected ovary showing oocytes in different developmental stages, the primary symbiont *Portiera* inside bacteriocyte cells in *red* and the secondary symbiont *Rickettsia* in *green*. (**f**) A portion of dissected midgut showing DAPI staining of the nuclei in *blue* and the secondary symbiont *Rickettsia* in *red. Green* represents FISH for localizing heat shock protein 70

Rickettsia and TYLCV in this study, in which high levels of the bacterium in the midgut resulted in higher virus concentrations in the filter chamber, a favored site for virus translocation along the transmission pathway, whereas low levels of *Rickettsia* in the midgut resulted in an even distribution of the virus.

Several recent projects have sequenced the genomes of some *B. tabaci* symbionts, including the primary symbiont, *Portiera* [111–113], and secondary symbionts, *Rickettsia* and *Hamiltonella* [100, 114], and *Cardinium* [115]. These projects have provided unique insights into the genome reduction of these symbionts, which has made them completely dependent on their hosts and, therefore, not culturable. Additionally, the genome sequences of these symbionts have shown the dependence of the insect host on amino acid biosynthetic pathways lacking in the insect genome, providing further insights on the mutual symbiosis between the insect and its symbiotic bacteria. This is more evident when examining the genome of *Portiera*, the primary symbiont, which has an extremely reduced genome, yet maintains biosynthetic pathways essential for the survival of the symbiont as well as the insect.

Several additional recent *B. tabaci* EST sequencing projects identified a total of 17,766 bacterial unigenes which were classified into 322 genera [(*Proteobacteria* (92 %), *Betaproteobacteria* (59 %), *Burkholderiales* (58 %), *Comamonadaceae* (50 %), and *Delftia* (43 %)] and revealed a highly diverse bacterial community which represents a high and rapid coevolution of insects and their symbionts [79]. Additional population surveys from Brazil using rRNA gene sequencing showed that *Hamiltonella* and *Rickettsia* are highly prevalent in all MEAM1 populations tested, while *Cardinium* was close to fixation in only three populations. Surprisingly, some MEAM1 individuals and one New World 2 population were infected with *Fritschea* [90]. Guo et al. [116] observed more females than males harboring *Hamiltonella*, and the results suggested that both the female-biased symbiont infections and female-biased TYLCV infections promote the rapid spread of TYLCV in China. Recently, Zchori-Fein et al. [97] modeled the variations in the symbiotic communities of about 2000 sweetpotato whitefly individuals and demonstrated facultative endosymbiotic combinations which were positively correlated with both distance from the equator and specificity of the genetic code of the insect host.

The reports cited here show that begomovirus translocation in the vector involves not only hypothesized insect proteins and receptors but also bacterial proteins, which could be acting directly or indirectly to influence the virus transmission. Future whitefly genomics research is expected to gain better understanding of the novel genes and mechanisms involved in virus-vector-symbiont relationships.

2.6 Studying the Interaction Between *B. tabaci* and Begomoviruses

B. tabaci is known to transmit a great diversity of plant viruses; among them, begomoviruses are the most devastating and cause serious viral diseases in agricultural cropping systems worldwide. The whitefly-begomovirus interactions depend on the virus, the biotype of the whitefly vector, and the endosymbionts harbored in the specific vector [68]. More than 80 % of the mono- and bipartite begomoviruses (both New and Old worlds) are transmitted by whiteflies in and around tropical and subtropical regions. Whitefly-transmitted begomoviruses are known to cause leaf curling, yellow mosaic, and yellow vein mosaic in several important crops and weeds. Selection pressure exerted by vectors of a particular geographical location is a crucial factor in the evolution of begomoviruses. Excellent reviews are available on the whitefly-transmitted viruses in New World and Old World and their interaction with insect vectors [21, 68, 117–122].

Begomoviruses are vectored by *B. tabaci* in a persistent and circulative manner (Fig. 2.3). The stylet of *B. tabaci* penetrates the plant epidermis and moves intracellularly through the parenchyma to reach the phloem, which is required for both virus acquisition and transmission. Whiteflies acquire virions during feeding from the phloem of an infected plant. The virions move through the alimentary canal into the whitefly midgut (Fig. 2.4), where they enter the hemolymph and transit to the salivary glands for transmission during the next feeding cycle [123] (Fig. 2.3).

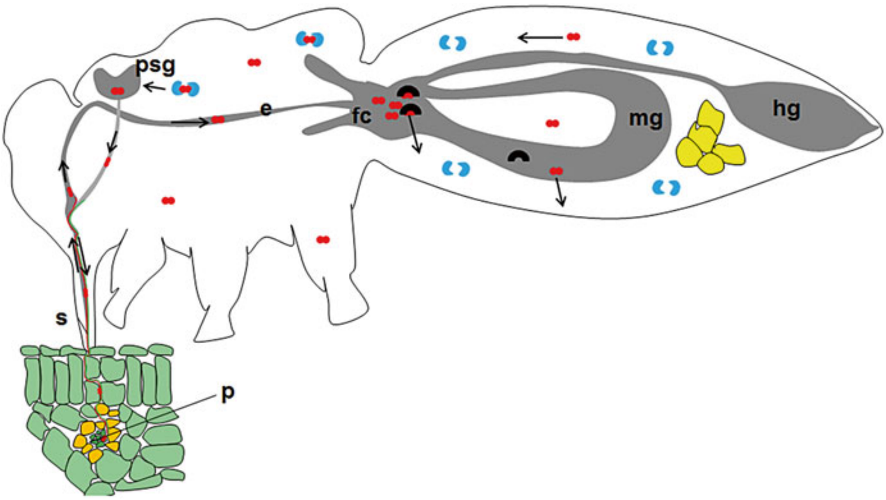


Fig. 2.3 Schematic representation of the circulative route of *tomato yellow leaf curl virus* (TYLCV) inside *B. tabaci.* The virus (*red particles*) is acquired from the plant phloem (*p*) through the stylet (*s*). The virus moves along the esophagus (*e*) and reaches the filter chamber (*fc*) in the midgut (*mg*) where the majority of the virus particles are translocated into the hemolymph. This translocation is aided, in part, by the heat shock protein 70 (*black particles*). In the hemolymph, *Hamiltonella*, a bacterial symbiont harbored within bacteriocyte cells (*yellow organs*), produces a GroEL protein (*blue particles*) that interacts with TYLCV and aids in its safe arrival to the primary salivary glands (*psg*). Once inside the salivary glands, the virus moves along the salivary duct to be injected into a new plant while the insect is feeding. *hg* hindgut

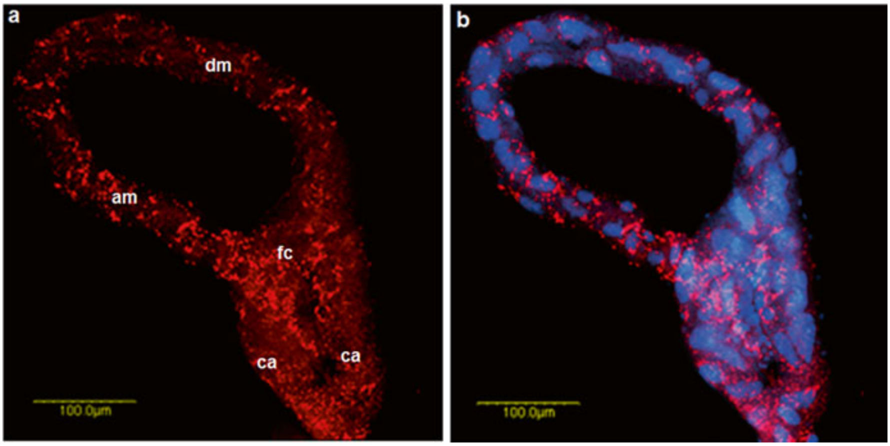


Fig. 2.4 Immunolocalization of TYLCV in the midgut of *B. tabaci* following acquisition of the virus from infected plants for several days. First antibody was used against the virus CP and secondary antibody attached to Cy3 dye (*red*). (**a**) Shows the virus in *red*, scattered along the midgut and concentrating in the filter chamber (*fc*). (**b**) The same midgut as in the *left panel* showing DAPI staining of the nuclei (*blue*). *am* ascending midgut, *dm* descending midgut, *ca* ceca, *CP* coat protein

The capsid protein plays a crucial role in the transmission of begomoviruses by *B. tabaci*. Evidence for a CP role in virus transmission emerged from following observations: (a) exchange of an *African cassava mosaic virus* (ACMV) CP gene with that of *beet curly top virus* (BCTV) altered the insect specificity of ACMV from whiteflies to leafhoppers [124]; (b) whiteflies have been shown to be unable to acquire coat protein mutants of geminiviruses that did not form capsids [125, 126]; and (c) replacing the CP of the non-transmissible begomovirus, *Abutilon mosaic virus* (AbMV), by that of a transmissible isolate of *Sida golden mosaic virus* (SiGMV), restored transmission by *B. tabaci* [127].

Site-directed mutagenesis in the CP domain was used to map the functional regions in *tomato yellow leaf curl Sardinia virus* (TYLCSV) which are involved in the transmission of the virus by *B. tabaci* [128, 129]. Briefly, a region between amino acids 129 and 152, including Q129, Q134, and D152, was found to be relevant for virion assembly, systemic infection, and transmission by the vector. In addition, it was shown that crossing the salivary gland barrier may not be sufficient for transmission. These results were confirmed in other bipartite begomoviruses, namely, *watermelon chlorotic stunt virus* (WmCSV) [130] and AbMV [131]. Using immunogold labeling studies, the CP mutants PNHD and PNHE were detected in *B. tabaci* salivary glands similar to the wild-type virus, but the mutant QDQD CP was not detected in salivary gland cells. The most interesting result was that although the mutants PNHD and PNHE were present in primary salivary gland, transmission did not occur, suggesting that their presence in salivary gland does not ensure transmission. Thus, molecular interactions with saliva might be necessary to maintain infectivity, which, perhaps, are absent from these mutants [129].

Further studies to unravel the importance of TYLCV CP in virus transmission showed its interaction with a member of the small heat shock protein family (BtHSP16), which was identified using a yeast two-hybrid system screen against TYLCSV CP [132]. Another study recently demonstrated that another heat shock protein, HSP70, interacts with the TYLCV CP in vivo and in vitro, and membrane feeding with anti-HSP70 antibodies resulted in an increase in TYLCV transmission. This result suggested that under normal conditions HSP70 restricts virus activity, thereby protecting the insect from deleterious effects of the circulating virus [133]. Interactions with these proteins may be necessary for refolding of the virion particle and facilitating its crossing the barrier between the midgut and the hemolymph or hemolymph and the salivary glands.

2.7 Interaction of *B. tabaci* with Plants Using Genomics Tools

Many plants have developed various chemical defenses to deter or combat insect pests. To date, some genomic studies were conducted for understanding the plant responses to *B. tabaci*. This has limited our understanding of the plant response to

hemipteran species in general. Hemipterans cause minimal cellular wounding, and although the duration of feeding can be very long, the molecular plant responses can be minimal; however, they can also be dramatic [134]. Salicylic acid (SA) and jasmonic acid (JA) are two important components in the signal transduction cascades that regulate plant defense responses against biotic stresses. JA affects plant resistance to necrotrophic pathogens and tissue-damaging insects as well as some phloem-feeding insects [135]. Other plant defense proteins such as proteinase inhibitors (PI) [136, 137], PI II [138–140], LOX genes [140], and others are inducible during insect feeding.

To understand the major plant defenses in responding to feeding by the whitefly, studies were conducted with the model plant *Arabidopsis thaliana* [141, 142]. Zarate et al. [141] reported that the whitefly induces SA defenses and suppresses effectual JA defenses. Supporting this study, using the Affymetrix ATH1 GeneChip to monitor the *A. thaliana* transcriptome, 700 transcripts were found to be upregulated and 556 downregulated by feeding of the silverleaf whitefly nymphs [142]. Closer examination of the regulation of secondary metabolite (glucosinolate) and defense pathway genes after feeding showed that responses were qualitatively and quantitatively different from chewing insects and aphids. Furthermore, the JA genes PDF1.2, PR4, and CORI3 were repressed and SA-regulated genes were induced after whitefly feeding. Cytological staining of whitefly-infected tissue showed that pathogen defenses, such as localized cell death and hydrogen accumulation, were not observed [142].

Few studies have shown that virus proteins can act as suppressors for plant defense systems which are directed to block or reduce defenses against insect feeding. Coinfection of the begomovirus TYLCCNV and its *tomato yellow leaf curl China virus* betasatellite (TYLCCNB) could repress the JA response but not the SA response [143]. Microarray analysis of the *A. thaliana* transcriptome in response to *cabbage leaf curl virus* (CabLCV) infection uncovered 5365 differentially expressed genes; data mining revealed that CabLCV triggers a pathogen response via the SA pathway and induces expression of genes involved in programmed cell death, genotoxic stress, and DNA repair [144]. Further, coinfection of plants by TYLCCNV and TYLCCNB increased expression of the whitefly vitellogenin gene [145]. Recently, Shi et al. [146] demonstrated that SA content was always higher in leaves infested with the viruliferous B biotype compared with the viruliferous Q biotype. Additional recent studies by Shi et al. [147, 148] showed that *B. tabaci* Q carrying TYLCV strongly suppressed host plant defenses.

2.8 Interactions of *B. tabaci* with Parasitoids and Pathogens Using Genomic Tools

Parasitoids that specifically attack *B. tabaci* have been used commercially for the biocontrol of whiteflies. Most whitefly parasitoids kill the insect by piercing the body contents and feeding on host fluids. Parasitoids from the genera *Encarsia*,

Eretmocerus (*Aphelinidae*), and *Amitus* (*Platygastridae*) can attack whiteflies [149, 150]. Recent studies identified *Encarsia sophia* and *Eretmocerus hayati* (*Hymenoptera*: *Aphelinidae*) as two key parasitoids of *B. tabaci* (MEAM1) [151, 152], and *En. formosa* [153], for which it was suggested that begomoviruses transmitted by the whitefly can manipulate the host suitability of a parasitoid and hence the parasitoid-host interactions.

Mahadav et al. [84] investigated the response of *B. tabaci* to parasitization by the wasp, *E. mundus*, using a whitefly cDNA-based microarray. The results clearly indicated that genes known to be part of defense pathways described in other insects are also involved in the response of *B. tabaci* to parasitization by *E. mundus*. Some of these responses include repression of a serine protease inhibitor (*serpin*) and induction of a melanization cascade. Quantitative real-time RT-PCR (qRT-PCR) and fluorescence in situ hybridization (FISH) analyses showed that the proliferation of *Rickettsia*, a facultative secondary symbiont, was strongly induced upon initiation of the parasitization process [84].

Recently, using RNA-seq technology, Zhang et al. [154] conducted a comprehensive investigation of the whitefly defense response to infection by *Pseudomonas aeruginosa*. Compared to uninfected whiteflies, 6- and 24-h post-infected whiteflies showed 1348 upregulated and 1888 downregulated genes that were differentially expressed. Functional analysis of these genes revealed that the mitogen-associated protein kinase (MAPK) pathway was activated after *P. aeruginosa* infection. Additional research included investigating the response of *B. tabaci* to the entomopathogenic fungus, *Beauveria bassiana*, using next-generation sequencing, which showed that conserved pathways such as the DNA repair and MAPK pathways are responsive to attack by this fungus [155].

2.9 Concluding Remarks and Future Perspective

B. tabaci is undoubtedly one of the major worldwide insect pests in agriculture. Despite many years of research, conventional chemical control methods for reducing the damage caused by this pest are still dominating. The past two decades have witnessed a dramatic increase in the research aimed toward understanding the interactions of *B. tabaci* with its environment and other organisms, including those that are part of its damage cycle such as plant viruses. However, those interactions also include plants, parasitoids, symbionts, and abiotic factors such as heat. The use of newly developed molecular and genomic tools will facilitate our understanding of the genetic makeup of the whitefly and will enable further development of novel tools for better means of controlling this insect pest.

References

1. Severson DW, Behura SK (2012) Mosquito genomics: progress and challenges. Annu Rev Entomol 57:143–166
2. Brown JK (1994) Current status of *Bemisia tabaci* as a plant pest and virus vector in agro-ecosystems worldwide. FAO Plant Prot Bull 42:3–32
3. Brown JK, Frohlich DR, Rosell RC (1995) The sweetpotato or silverleaf whiteflies: biotypes of *Bemisia tabaci* or a species complex? Annu Rev Entomol 40:511–534
4. Costa HS, Brown JK, Sivasupramaniam S, Bird J (1993) Regional distribution, insecticide resistance, and reciprocal crosses between the 'A' and 'B' biotypes of *Bemisia tabaci*. Insect Sci Appl 14:255–266
5. Gennadius P (1889) Disease of tobacco plantations in the Trikonia. The aleyrodid of tobacco. Ellenike Georgia 5:1–3
6. Quaintance AL (1900) Contribution towards a monograph of the American aleurodidae. US Department of Agriculture. Technica series. Bur Entomol 8:9–64
7. Russell LM (1957) Synonyms of *Bemisia tabaci* (Gennadius) (Homoptera, Aleyrodidae). Bull Brooklyn Entomol Soc 52:122–123
8. Perring TM (2001) The *Bemisia tabaci* species complex. Crop Prot 20:725–737
9. Oliveira MRV, Henneberry TJ, Anderson P (2001) History, current status, and collaborative research projects for *Bemisia tabaci*. Crop Prot 20:709–723
10. Dinsdale A, Cook L, Riginos C, Buckley YM, De Barro P (2010) Refined global analysis of *Bemisia tabaci* (Hemiptera: Sternorrhyncha: Aleyrodidae: Aleyrodidae) mitochondrial cytochrome oxidase 1 to identify species level genetic boundaries. Ann Entomol Soc Am 103:196–208
11. De Barro PJ, Liu SS, Boykin L, Dinsdale A (2011) *Bemisia tabaci*: a statement of species status. Annu Rev Entomol 56:1–19
12. Firdaus S, Vosman B, Hidayati N, Supena E, Visser RGF, van Heusden AW (2013) The *Bemisia tabaci* species complex: additions from different parts of the world. Insect Sci 20:723–733
13. Boykin LM, Bell CD, Evans G, Small I, De Barro PJ (2013) Is agriculture driving the diversification of the *Bemisia tabaci* species complex (Hemiptera: Sternorrhyncha: Aleyrodidae)?: dating, diversification and biogeographic evidence revealed. BMC Evol Biol 13:228
14. Misra CS, Lamba SK (1929) The cotton whitefly (*Bemisia gossypiperda* n. sp.). Bull Agric Res Inst Pusa 196:1–7
15. Husain MA, Trehan KN (1933) Observations on the life history, bionomics and control of the whitefly of cotton (*Bemisia gossypierda*, n. sp.). Bull Agric Res Inst Pusa 196:7
16. Martin JH, Mifsud D, Rapisarda C (2000) The whiteflies (Hemiptera: Aleyrodidae) of Europe and the Mediterranean Basin. Bull Entomol Res 90:407–448
17. Jones DR (2003) Plant viruses transmitted by whiteflies. Eur J Plant Pathol 109:195–219
18. Mugiira RB, Liu SS, Zhou X (2008) *Tomato yellow leaf curl virus and Tomato leaf curl Taiwan virus* invade south-east coast of China. J Phytopathol 156:217–221
19. Riley DG, Palumbo JC (1995) Interaction of silverleaf whitefly (Homoptera: Aleyrodidae) with cantaloupe yield. J Econ Entomol 88:1726–1732
20. Anderson PK, Morales FJ (2005) Whitefly and whitefly-borne viruses in the tropics: building a knowledge base for global action. CIAT, Cali, p 351
21. Navas-Castillo J, Fiallo-Olive E, Sanchez-Campos S (2011) Emerging virus diseases transmitted by whiteflies. Ann Rev Phytopathol 49:219–248
22. Brown JK (2000) Molecular markers for the identification and global tracking of whitefly vector-Begomovirus complexes. Virus Res 71:233–260
23. Liu SS, De Barro PJ, Xu J, Luan JB, Zang LS, Ruan YM, Wan FH (2007) Asymmetric mating interactions drive widespread invasion and displacement in a whitefly. Science 318:1769–1772

24. Jiu M, Zhou X-P, Tong L, Xu J, Yang X, Wan FH, Liu SS (2007) Vector-virus mutualism accelerates population increase of an invasive whitefly. PLoS One 2:e182
25. Crowder DW, Horowitz AR, De Barro PJ (2010) Mating behaviour, life history and adaptation to insecticides determine species exclusion between whiteflies. J Anim Ecol 79:563–570
26. Gorman K, Slater R, Blande J, Clarke A, Wren J, McCaffery A, Denholm I (2010) Cross-resistance relationships between neonicotinoids and pymetrozine in *Bemisia tabaci* (Hemiptera: Aleyrodidae). Pest Manag Sci 66:1186–1190
27. Brown JK, Czosnek H (2002) Whitefly transmission of plant viruses. In: Plumb RT (ed) Advances in botanical research, vol 36. Academic, New York
28. Bird J (1957) A whitefly transmitted mosaic of *Jatropha gossypifolia*. Technical paper. University of Puerto Rico, Agricultural Experiment Station 22:1–35
29. Bird J, Maramorosch K (1978) Viruses and virus diseases associated with whiteflies. Adv Virus Res 22:55–110
30. Costa HS, Russel M (1975) Failure of *Bemisia tabaci* to breed on cassava plants in Brazil (Homoptera, Aleyrodidae). Cienia Cult 27:388–390
31. Perring TM, Cooper A, Kazmer DJ, Shields C, Sheilds J (1991) New strain of sweet potato whitefly invades California vegetables. Calif Agric 45:10–12
32. Bedford ID, Briddon RW, Brown JK, Rossel RC, Markham PG (1994) Geminivirus transmission and biological characterisation of *Bemisia tabaci* (Genn) biotypes from different geographic regions. Ann Appl Biol 125:311–325
33. Bellows TS, Perring TM, Gill RJ, Headrich DH (1994) Description of a species of *Bemisia tabaci* (Homoptera: Aleyrodidae). Ann Entomol Soc Am 87:195–206
34. Perring TM, Cooper AD, Rodrigues RJ, Farrar CA, Bellows TSJ (1993) Identification of a whitefly species by genomic and behavioural studies. Science 259:74–77
35. Gawel NJ, Bartlett AC (1993) Characterization of differences between whiteflies using RAPD-PCR. Insect Mol Biol 2:33–38
36. Chu D, Zhang YJ, Cong B, Xu BY, Wu QJ (2004) Developing sequence characterized amplified regions (SCARs) to identify *Bemisia tabaci* and *Trialeurodes vaporariorum*. Plant Prot 30:27–30
37. Khasdan V, Levin I, Rosner A, Morin S, Kontsedalov S et al (2005) DNA markers for identifying biotypes B and Q of *Bemisia tabaci* (Hemiptera: Aleyrodidae) and studying population dynamics. Bull Entomol Res 95:605–613
38. Zang LS, Chen WQ, Liu SS (2006) Comparison of performance on different host plants between the B biotype and a non-B biotype of *Bemisia tabaci* from Zhejiang, China. Entomol Exp Appl 121:221–227
39. Boukhatem N, Jdaini S, Mukovski Y, Jacquemin JM, Bouali A (2007) Identification of *Bemisia tabaci* (Gennadius) (Homoptera: Aleyrodidae) based on RAPD and design of two SCAR markers. J Biol Res (Thessaloniki) 8:167–176
40. Ko CC, Hung YC, Wang CH (2007) Sequence characterized amplified region markers for identifying biotypes of *Bemisia tabaci* (Hem., Aleyrodidae). J Appl Entomol 131:542–547
41. Shatters RG Jr, Powell CA, Boykin LM, He LS, McKenzie CL (2009) Improved DNA barcoding method for *Bemisia tabaci* and related Aleyrodidae: development of universal and *Bemisia tabaci* biotype-specific mitochondrial cytochrome c oxidase chain reaction primers. J Econ Entomol 102:750–758
42. Ma DY, Li XC, Dennehy TJ, Lei CL, Wang M et al (2009) Utility of mtCO1 polymerase chain reaction-restriction fragment length polymorphism in differentiating between Q and B whitefly *Bemisia tabaci* biotypes. Insect Sci Appl 16:107–114
43. Cervera MT, Cabezas JA, Simon B, Martinez-Zapater JM, Beitia F, Cenis JL (2000) Genetic relationships among biotypes of *Bemisia tabaci* Hemiptera, Aleyrodidae based on AFLP analysis. Bull Entomol Res 90:391–396
44. Frohlich DR, Torres-Jerez I, Bedford ID, Markham PG, Brown JK (1999) A phylogeographical analysis of the *Bemisia tabaci* species complex based on mitochondrial DNA markers. Mol Ecol 8:1683–1691

45. Hu J, De Barro P, Zhao H, Wang J, Nardi F (2011) An extensive field survey combined with a phylogenetic analysis reveals rapid and widespread invasion of two alien whiteflies in China. PLoS One 6:e16061
46. Alemandri V, De Barro PJ, Bejerman N, Argüello Caro EB, Dumón AD, Mattio MF, Rodriguez SM, Truoli G (2012) Species within the *Bemisia tabaci* (Hemiptera: *Aleyrodidae*) complex in soybean and bean crops in Argentina. J Econ Entomol 105:48–53
47. Boykin LM, Armstrong KF, Kubatko L, De Barro PJ (2012) Species delimitation and global biosecurity. Evol Bioinform 8:1–37
48. Liu SS, Colvin J, De Barro PJ (2012) Species concepts as applied to the whitefly *Bemisia tabaci* systematics: how many species are there? J Integr Agric 11:176–186
49. Lee W, Park J, Lee G, Lee S, Akimoto S (2013) Taxonomic status of the *Bemisia tabaci* complex (Hemiptera: Aleyrodidae) and reassessment of the number of its constituent species. PLoS One 8:e63817
50. De Barro PJ, Driver F, Trueman JWH, Curran J (2000) Phylogenetic relationship of world populations of *Bemisia tabaci* (Gennadius) using ribosomal ITS1. Mol Phylogenet Evol 16:29–36
51. Wu X, Li Z, Hu D, Shen Z (2003) Identification of Chinese populations of *Bemisia tabaci* Gennadius by analyzing ITS1 sequence. Prog Nat Sci 13:276–281
52. Abdullahi I, Winter S, Atirim GI, Thottappilly G (2003) Molecular characterization of whitefly, *Bemisia tabaci* Hemiptera, Aleyrodidae populations infesting cassava. Bull Entomol Res 93:97–106
53. De Barro PJ (2005) Genetic structure of the whitefly *Bemisia tabaci* in the Asia-Pacific region revealed using microsatellite markers. Mol Ecol 14:3695–3718
54. Li ZX, Lin HZ, Guo XP (2007) Prevalence of *Wolbachia* infection in *Bemisia tabaci*. Curr Microbiol 54:467–471
55. De Barro PJ, Scott KD, Graham GC, Lange CL, Schutze MK (2003) Isolation and characterization of microsatellite loci in *Bemisia tabaci*. Mol Ecol Notes 3:40–43
56. Tsagkarakou A, Roditakis N (2003) Isolation and characterization of microsatellite loci in *Bemisia tabaci* (Hemiptera: Aleyrodidae). Mol Ecol Notes 3:196–198
57. De Barro PJ, Trueman JWH, Frohlich DR (2005) *Bemisia argentifolii* is a race of *B. tabaci* (Hemiptera: Aleyrodidae): the molecular genetic differentiation of *B. tabaci* populations around the world. Bull Entomol Res 95:193–203
58. De Barro PJ, Hidayat SH, Frohlich D, Subandiyah S, Ueda S (2008) A virus and its vector, pepper yellow leaf curl virus and *Bemisia tabaci*, two new invaders of Indonesia. Biol Invasions 10:411–433
59. Delatte H, David P, Granier M, Lett JM, Goldbach R, Peterschmitt M, Reynaud B (2006) Microsatellites reveal extensive geographical, ecological and genetic contacts between invasive and indigenous whitefly biotypes in an insular environment. Genet Res 87:109–124
60. Delatte H, Reynaud B, Granier M, Thornary L, Lett JM, Goldbach R, Peterschmitt M (2005) A new silverleaf-inducing biotype Ms of *Bemisia tabaci* (Hemiptera: Aleyrodidae) indigenous to the islands of the southwest Indian Ocean. Bull Entomol Res 95:29–35
61. Dalmon A, Halkett F, Granier M, Delatte H, Peterschmitt M (2008) Genetic structure of the Invasive pest *Bemisia tabaci*: evidence of limited but persistent genetic differentiation in glasshouse populations. Heredity 100:316–325
62. Tsagkarakou A, Tsigenopoulos CS, Gorman K, Lagnel J, Bedford ID (2007) Biotype status and genetic polymorphism of the whitefly *Bemisia tabaci* (Hemiptera: Aleyrodidae) in Greece: mitochondrial DNA and microsatellites. Bull Entomol Res 97:29–40
63. Gauthier N, Dalleau-Clouet C, Bouvret M-E (2008) Twelve new polymorphic microsatellite loci and PCR multiplexing in the whitefly, *Bemisia tabaci*. Mol Ecol Resour 8:1004–1007
64. Valle G, Lourençao A, Zucchi M, Pinheiro J (2012) Low polymorphism revealed in new microsatellite markers for *Bemisia tabaci* (Hemiptera: Aleyrodidae). Genet Mol Res 11:3899–3903
65. Wang HL, Yang J, Boykin LM, Zhao QY, Wang YJ, Liu SS, Wang XW (2014) Developing conversed microsatellite markers and their implications in evolutionary analysis of the *Bemisia tabaci* complex. Sci Rep 4:6351

66. Tay WT, Evans GA, Boykin LM, De Barro PJ (2012) Will the real *Bemisia tabaci* please stand up? PLoS ONE 7:e50550
67. Brown JK (2007) The *Bemisia tabaci* complex: genetic and phenotypic variation and relevance to TYLCV-vector interactions. In: Czosnek H (ed) Tomato yellow leaf curl virus disease. Springer, Dordrecht, pp 25–56
68. Ghanim M (2014) A review of the mechanisms and components that determine the transmission efficiency of Tomato yellow leaf curl virus (*Geminiviridae*; *Begomovirus*) by its whitefly vector. Virus Res 186:47–54
69. Czosnek H, Brown J (2010) The whitefly genome – white paper: proposal to sequence multiple genomes of *Bemisia tabaci*. In: Stansly PA, Naranjo SE (eds) *Bemisia*: bionomics and management of a global pest. Springer, Dordrecht, pp 503–532, 540 pages
70. Brown JK, Lambert GM, Ghanim M, Czosnek H, Galbraith DW (2005) Nuclear DNA content of the whitefly *Bemisia tabaci* (Genn.) (Aleyrodidae: Homoptera/Hemiptera) estimated by flow cytometry. Bull Entomol Res 95:309–312
71. Leshkowitz D, Gazit S, Reuveni E, Ghanim M, Czosnek H, McKenzie C, Shatters RG Jr, Brown JK (2006) Whitefly (*Bemisia tabaci*) genome project: analysis of sequenced clones from egg, instar, and adult (viruliferous and non-viruliferous) cDNA libraries. BMC Genomics 7:79
72. Ghanim M, Kontsedalov S (2007) Gene expression in pyriproxyfen-resistant *Bemisia tabaci* Q biotype. Pest Manag Sci 63:776–783
73. Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS (2010) *De novo* characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11:400
74. Luan JB, Li JM, Varela N, Wang YL, Li FF, Bao YY, Zhang CX, Liu SS, Wang XW (2011) Global analysis of the transcriptional response of whitefly to tomato yellow leaf curl china virus reveals their relationship of coevolved adaptations. J Virol 85:3330–3340
75. Rubinstein G, Czosnek H (1997) Long-term association of tomato yellow leaf curl virus with its whitefly vector *Bemisia tabaci*: effect on the insect transmission capacity, longevity and fecundity. J Gen Virol 78:2683–2689
76. Su YL, Li JM, Li M, Luan JB, Ye XD, Wang XW, Liu SS (2012) Transcriptomic analysis of the salivary glands of an invasive whitefly. PLoS One 7:e39303
77. Wang XW, Luan JB, Li JM, Su YL, Xia J, Liu SS (2011) Transcriptome analysis and comparison reveal divergence between two invasive whitefly cryptic species. BMC Genomics 12:458
78. Wang XW, Zhao QY, Luan JB, Wang YJ, Yan GH, Liu SS (2012) Analysis of a native whitefly transcriptome and its sequence divergence with two invasive whitefly species. BMC Genomics 13:529
79. Xie W, Meng Q, Wu Q, Wang S, Yang X, Yang N, Li R, Jiao X, Pan H, Liu B, Su Q, Xu B, Hu S, Zhou X, Zhang Y (2012) Pyrosequencing the *Bemisia tabaci* transcriptome reveals a highly diverse bacterial community and a robust system for insecticide resistance. PLoS One 7:e35181
80. Wang HL, Yang J, Boykin LM, Zhao QY, Li Q, Wang XW, Liu SS (2013) The characteristics and expression profiles of the mitochondrial genome for the Mediterranean species of the *Bemisia tabaci* complex. BMC Genomics 14:401
81. Xie W, Yang X, Wang S, Wu Q, Yang N, Li R, Jiao X, Pan H, Liu B, Feng Y, Xu B, Zhou X, Zhang Y (2012) Gene expression profiling in the thiamethoxam resistant and susceptible B-biotype sweetpotato whitefly, *Bemisia tabaci*. J Insect Sci 12:46
82. Yang N, Xie W, Yang X, Wang S, Wu Q, Li R, Pan H, Liu B, Shi X, Fang Y, Xu B, Zhou X, Zhang Y (2014) Transcriptomic and proteomic responses of sweetpotato whitefly, *Bemisia tabaci*, to Thiamethoxam. PLoS One 8:e61820
83. Ye XD, Su YL, Zhao QY, Xia WQ, Liu SS, Wang XW (2014) Transcriptomic analyses reveal the adaptive features and biological differences of guts from two invasive whitefly species. BMC Genomics 15:370
84. Mahadav A, Gerling D, Gottlieb Y, Czosnek H, Ghanim M (2008) Gene expression in the whitefly *Bemisia tabaci* pupae in response to parasitization by the wasp *Eretmocerus mundus*. BMC Genomics 9:342

85. Chiel E, Gottlieb Y, Zchori-Fein E, Mozes-Daube N, Katzir N, Inbar M, Ghanim M (2007) Biotype-dependent secondary symbiont communities in sympatric populations *of Bemisia tabaci*. Bull Entomol Res 97:407–413
86. Skaljac M, Zanic K, Goreta-Ban S, Kontsedalov S, Ghanim M (2010) Co-infection and localization of secondary symbionts in two whitefly species. Isr J Plant Sci 58:103–111
87. Skaljac M, Zanic K, Hrncic S, Radonjic S, Perovic T, Ghanim M (2013) Diversity and localization of bacterial symbionts in three whitefly species (Hemiptera: Aleyrodidae) from the east coast of the Adriatic Sea. Bull Entomol Res 103:48–59
88. Gueguen G, Vavre F, Gnankine O, Peterschmitt M, Charif D, Chiel E, Gottlieb Y, Ghanim M, Zchori-Fein E, Fleury F (2010) Endosymbiont metacommunities, mtDNA diversity and the evolution of the *Bemisia tabaci* (Hemiptera: Aleyrodidae) species complex. Mol Ecol 19:4365–4376
89. Henry LM, Peccoud J, Simon JC, Hadfield JD, Maiden MJ, Ferrari J, Godfray HC (2013) Horizontally transmitted symbionts and host colonization of ecological niches. Curr Biol 23:1713–1717
90. Marubayashi JM, Kliot A, Yuki VA, Marques-Rezende JA, Krause-Sakate R, Pavan MA, Ghanim M (2014) Diversity and localization of bacterial endosymbionts from whitefly species collected in Brazil. PLoS One 9:e108363
91. Gottlieb Y, Ghanim M, Gueguen G, Kontsedalov S, Vavre F, Fleury F, Zchori-Fein E (2008) Inherited intracellular ecosystem: symbiotic bacteria share bacteriocytes in whiteflies. FASEB J 22:2591–2599
92. Chu D, Gao CS, De Barro P, Zhang YJ, Wan FH, Khan IA (2011) Further insights into the strange role of bacterial endosymbionts in whitefly, *Bemisia tabaci*: comparision of secondary symbionts from biotypes B and Q in China. Bull Entomol Res 101:477–486
93. Liu S, Chougule NP, Vijayendran D, Bonning BC (2012) Deep sequencing of the transcriptomes of soybean aphid and associated endosymbionts. PLoS One 7:e45161
94. Bing XL, Ruan YM, Rao Q, Wang XW, Liu SS (2013) Diversity of secondary endosymbionts among different putative species of the whitefly *Bemisia tabaci*. Insect Sci 20:194–206
95. Bing XL, Xia WQ, Gui JD, Yan GH, Wang XW, Liu SS (2014) Diversity and evolution of the *Wolbachia* endosymbionts of *Bemisia* (Hemiptera: Aleyrodidae) whiteflies. Ecol Evol 4:2714–2737
96. Helene D, Remy B, Nathalie B, Anne-Laure G, Traore RS, Jean-Michel L, Bernard R (2014) Species and endosymbiont diversity of *Bemisia tabaci* (Homoptera: Aleyrodidae) on vegetable crops in Senegal. Insect Sci. doi:10.1111/1744-7917.12134
97. Zchori-Fein E, Lahav T, Freilich S (2014) Variations in the identity and complexity of endosymbiont combinations in whitefly hosts. Front Microbiol 5:310
98. Thao ML, Baumann P (2004) Evolutionary relationships of primary prokaryotic endosymbionts of whiteflies and their hosts. Appl Environ Microbiol 70:3401–3406
99. Xue X, Li SJ, Ahmed MZ, De Barro PJ, Ren SX, Qiu BL (2012) Inactivation of *Wolbachia* reveals its biological roles in whitefly host. PLoS One 7:e48148
100. Rao Q, Wang S, Su YL, Bing XL, Liu SS, Wang XW (2012) Draft genome sequence of "Candidatus Hamiltonella defensa", an endosymbiont of the whitefly *Bemisia tabaci*. J Bacteriol 194:3558
101. Su Q, Xie W, Wang S, Wu Q, Liu B, Fang Y, Xu B, Zhang Y (2014) The endosymbiont *Hamiltonella* increases the growth rate of its host *Bemisia tabaci* during periods of nutritional stress. PLoS One 9:e89002
102. Brumin M, Kontsedalov S, Ghanim M (2011) *Rickettsia* influences thermotolerance in the whitefly *Bemisia tabaci* B biotype. Insect Sci 18:57–66
103. Kontsedalov S, Zchori-Fein E, Chiel E, Gottlieb Y, Inbar M, Ghanim M (2008) The presence of *Rickettsia* is associated with increased susceptibility of *Bemisia tabaci* (Homoptera: Aleyrodidae) to insecticides. Pest Manag Sci 64:789–792
104. Kontsedalov S, Gottlieb Y, Ishaaya I, Nauen R, Horowitz AR, Ghanim M (2009) Toxicity of spiromesifen to the developmental stages of *Bemisia tabaci* biotype B. Pest Manag Sci 65:5–13

105. Himler AG, Adachi-Hagimori T, Bergen JE, Kozuch A, Kelly SE, Tabashnik BE, Chiel E, Duckworth VE, Dennehy TJ, Zchori-Fein E, Hunter MS (2011) Rapid spread of a bacterial symbiont in an invasive whitefly is driven by fitness benefits and female bias. Science 332:254–256
106. Morin S, Ghanim M, Zeidan M, Czosnek H, Verbeek M, van den Heuvel JF (1999) A GroEL homologue from endosymbiotic bacteria of the whitefly *Bemisia tabaci* is implicated in the circulative transmission of tomato yellow leaf curl virus. Virology 256:75–84
107. Morin S, Ghanim M, Sobol I, Czosnek H (2000) The GroEL protein of the whitefly *Bemisia tabaci* interacts with the coat protein of transmissible and nontransmissible begomoviruses in the yeast two-hybrid system. Virology 276:404–416
108. Gottlieb Y, Zchori-Fein E, Mozes Daube N, Kontsedalov S, Skaljac M, Brumin M, Sobol I, Czosnek H, Vavre F, Fleury F, Ghanim M (2010) The transmission efficiency of *Tomato yellow leaf curl virus* by the whitefly *Bemisia tabaci* is correlated with the presence of a specific symbiotic bacterium species. J Virol 84:9310–9317
109. Rana VS, Singh ST, Priya NG, Kumar J, Rajagopal R (2012) *Arsenophonus* GroEL interacts with CLCuV and is localized in midgut and salivary gland of whitefly *B. tabaci*. PLoS One 7:e42168
110. Kliot A, Cilia M, Czosnek H, Ghanim M (2014) Implication of the Bacterial Endosymbiont *Rickettsia* spp. In interactions of the whitefly *Bemisia tabaci* with *Tomato yellow leaf curl virus*. Virology 88:5652–5660
111. Sloan DB, Moran NA (2012) Endosymbiotic bacteria as a source of carotenoids in whiteflies. Biol Lett 8:986–989
112. Jiang ZF, Xia FF, Johnson KW, Bartom E, Tuteja JH, Stevens R, Grossman RL, Brumin M, White KP, Ghanim M (2012) Genome sequences of the primary endosymbiont *Candidatus* Portiera aleyrodidarum in the whitefly *Bemisia tabaci* B and Q biotypes. J Bacteriol 194:6678–6679
113. Santos-Garcia D, Farnier PA, Beitia F, Zchori-Fein E, Vavre F, Mouton L, Moya A, Latorre A, Silva FJ (2012) Complete genome sequence of "*Candidatus* Portiera aleyrodidarum" BT-QVLC, an obligate symbiont that supplies amino acids and carotenoids to *Bemisia tabaci*. J Bacteriol 194:6654–6655
114. Rao Q, Wang S, Zhu DT, Wang XW, Liu SS (2012) Draft genome sequence of *Rickettsia* sp. strain MEAM1, isolated from the whitefly *Bemisia tabaci*. J Bacteriol 194:4741–4742
115. Santos-Garcia D, Latorre A, Moya A, Gibbs G, Hartung V, Dettner K, Kuechler SM, Silva FJ (2014) Small but powerful, the primary endosymbiont of moss bugs, *Candidatus* Evansia muelleri, holds a reduced genome with large biosynthetic capabilities. Genome Biol Evol 6:1875–1893
116. Guo H, Qu Y, Liu X, Zhong W, Fang J (2014) Female-biased symbionts and tomato yellow leaf curl virus infections in *Bemisia tabaci*. PLoS One 9:e84538
117. Morales FJ, Anderson PK (2001) The emergence and dissemination of whitefly-transmitted geminiviruses in Lain America. Arch Virol 146:415–441
118. Mansoor S, Briddon RW, Bull SE, Bedford ID, Bashir A, Hussain M, Saeed Zafar MY, Malik KA, Fauquet C, Markham PG (2003) Cotton leaf curl disease is associated with multiple monopartite begomoviruses supported by a single DNA b. Arch Virol 148:1969–1986
119. Varma A, Malathi VG (2003) Emerging geminivirus problems: a serious threat to crop production. Ann Appl Biol 142:145–164
120. Fargette D, Konate G, Fauquet C, Muller E, Peterschmitt M, Thresh JM (2006) Molecular ecology and emergence of tropical plant viruses. Annu Rev Phytopathol 44:235–260
121. Rojas MR, Gilbertson RL (2008) Emerging plant viruses: a diversity of mechanisms and opportunities. In: Roossinck MJ (ed) Plant virus evolution. Springer, Berlin/Heidelberg, pp 27–52
122. Hogenhout SA, Ammar ED, Whitfield AE, Redinbaugh MG (2008) Insect vector interactions with persistently transmitted viruses. Annu Rev Phytopathol 46:327–359

123. Ghanim M, Morin S, Czosnek H (2001) Rate of *Tomato yellow leaf curl virus* (TYLCV) translocation in the circulative transmission pathway of its vector, the whitefly *Bemisia tabaci*. Phytopathology 91:188–196
124. Briddon RW, Pinner MS, Stanley J, Markham PG (1990) Geminivirus coat protein gene replacement alters insect specificity. Virology 177:85–94
125. Azzam O, Frazer J, De La Rosa D, Beaver JS, Ahlquist P, Maxwell DP (1994) Whitefly transmission and efficient ssDNA accumulation of bean golden mosaic geminivirus require functional coat protein. Virology 204:289–296
126. Liu S, Bedford ID, Briddon RW, Markham PG (1997) Efficient whitefly transmission of bipartite geminiviruses requires both genomic components. J Gen Virol 78:1791–1794
127. Höfer P, Bedford ID, Markham PG, Jeske H, Frischmuth T (1997) Coat protein gene replacement results in whitefly transmission of an insect nontransmissible geminivirus isolate. Virology 236:288–295
128. Noris E, Vaira AM, Caciagli P, Masenga V, Gronenborn B, Accotto GP (1998) Amino acids in the capsid protein of tomato yellow leaf curl virus that are crucial for systemic infection, particle formation, and insect transmission. J Virol 72:10050–10057
129. Caciagli P, Medina Piles V, Marian D, Vecchiati M, Masenga V, Mason G, Falcioni T, Noris E (2009) Virion stability is important for the circulative transmission of *Tomato yellow leaf curl Sardinia virus* by *Bemisia tabaci*, but virion access to salivary glands does not guarantee transmissibility. J Virol 83:5784–5795
130. Kheyr-Pour A, Bananej K, Dafalla GA, Caciagli P, Noris E, Ahoonmanesh A, Lecoq H, Gronenborn B (2000) *Watermelon chlorotic stunt virus* from the Sudan and Iran: sequence comparisons and identification of a whitefly-transmission determinant. Phytopathology 90:629–635
131. Höhnle M, Höfer P, Bedford ID, Briddon RW, Markham PG, Frischmuth T (2001) Exchange of three amino acids in the coat protein results in efficient whitefly transmission of a nontransmissible *Abutilon mosaic virus* isolate. Virology 290:164–171
132. Ohnesorge S, Bejarano ER (2009) Begomovirus coat protein interacts with a small heat-shock protein of its transmission vector (*Bemisia tabaci*). Insect Mol Biol 18:693–703
133. Gotz M, Popovski S, Kollenberg M, Gorovits R, Brown JK, Cicero J, Czosnek H, Winter S, Ghanim M (2012) Implication of *Bemisia tabaci* heat shock protein 70 in begomovirus-whitefly interactions. J Virol 86:13241–13252
134. Walling LL (2000) The myriad plant responses to herbivores. J Plant Growth Regul 19:195–216
135. Pieterse CMJ, Dicke M (2007) Plant interactions with microbes and insects: from molecular mechanisms to ecology. Trends Plant Sci 12:564–568
136. Green TR, Ryan CA (1972) Wound-induced proteinase inhibitor in plant leaves: a possible defense mechanism against insects. Science 175:776–777
137. Thaler JS, Karban R, Ullman DE, Boege K, Bostock RM (2002) Cross-talk between jasmonate and salicylate plant defense pathways: effects on several plant parasites. Oecologia 131:227–235
138. Ryan CA (1990) Protease inhibitors in plants: genes for improving defenses against insects and pathogens. Annu Rev Phytopathol 28:425–449
139. Zhang HY, Xie XZ, Xu YZ, Wu NH (2004) Isolation and functional assessment of a tomato proteinase inhibitor II gene. Plant Physiol Biochem 42:437–444
140. Moran PJ, Thompson GA (2001) Molecular responses to aphid feeding in Arabidopsis in relation to plant defense pathways. Plant Physiol 125:1074–1085
141. Zarate SI, Kempema LA, Walling LL (2007) Silverleaf whitefly induces salicylic acid defenses and suppresses effectual jasmonic acid defenses. Plant Physiol 143:866–875
142. Kempema LA, Cui X, Holzer FM, Walling LL (2007) Arabidopsis transcriptome changes in response to phloem-feeding silverleaf whitefly nymphs. Similarities and distinctions in responses to aphids. Plant Physiol 143:849–865

143. Zhang T, Luan JB, Qi JF, Huang CJ, Li M, Zhou XP, Liu SS (2012) Begomovirus-whitefly mutualism is achieved through repression of plant defenses by a virus pathogenicity factor. Mol Ecol 21:1294–1304
144. Ascencio-Ibanez JT, Sozzani R, Lee TJ, Chu TM, Wolfinger RD, Cella R, Hanley-Bowdoin L (2008) Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection. Plant Physiol 148:436–454
145. Guo JY, Dong SZ, Yang X, Cheng L, Wan FH et al (2012) Enhanced vitellogenesis in a whitefly via feeding on a begomovirus-infected plant. PLoS One 7:e43567
146. Shi X, Pan H, Xie W, Wu Q, Wang S, Liu Y, Fang Y, Chen G, Gao X, Zhang Y (2013) Plant virus differentially alters the plant's defense response to its closely related vectors. PLoS One 8:e83520
147. Shi X, Pan H, Xie W, Jiao X, Fang Y, Chen G, Yang X, Wu Q, Wang S, Zhang Y (2014) Three-way interactions between the tomato plant, tomato yellow leaf curl virus, and the whitefly *Bemisia tabaci* facilitate virus spread. J Econ Entomol 107:920–926
148. Shi X, Pan H, Zhang H, Jiao X, Xie W, Wu Q, Wang S, Fang Y, Chen G, Zhou X, Zhang Y (2014) *Bemisia tabaci* Q carrying tomato yellow leaf curl virus strongly suppresses host plant defenses. Sci Rep 4:5230
149. Gerling D, Alomar O, Arno J (2001) Biological control of *Bemisia tabaci* using predators and parasitoids. Crop Prot 20:779–799
150. Huang J, Zheng QH, Fu JW, Huang PY, Gu DX (2000) Investigation and identification of the whitefly parasitoids (Hymenoptera: Aphelinidae, Platygasteridae). Entomol J East China 9:29–33
151. Yang NW, Wan FH (2011) Host suitability of different instars of *Bemisia tabaci* biotype B for the parasitoid *Eretmocerus hayati*. Biol Control 2:313–317
152. Xu HY, Yang NW, Wan FH (2013) Competitive interactions between parasitoids provide new insight into host suppression. PLoS One 8:e82003
153. Liu X, Xiang W, Jiao X, Zhang Y, Xie W, Wu Q, Zhou X, Wang S (2014) Effects of plant virus and its insect vector on -*Encarsia formosa*, a biocontrol agent of whiteflies. Sci Rep 4:5926
154. Zhang CR, Zhang S, Xia J, Li FF, Xia WQ, Liu SS, Wang XW (2014) The immune strategy and stress response of the Mediterranean species of the *Bemisia tabaci* complex to an orally delivered bacterial pathogen. PLoS One 9:e94477
155. Xia J, Zhang C, Zhang S, Li F, Feng M et al (2013) Analysis of whitefly transcriptional responses to *Beauveria bassiana* infection reveals new insights into insect-fungus interactions. PLoS One 8:e68185

Chapter 3
Updating Genomic Data of Lepidoptera

Carmen Pozo, Blanca Prado, and América Nitxin Castañeda-Sortibrán

Abstract Among the insects, lepidopterans form the second most diverse group, with over 155,000 described species. Research on Lepidoptera has a long tradition in several fields, including taxonomy, phylogeny, ecology, population genetics, evolutionary biology, speciation, physiology, development and gene regulation, host–plant and insect–parasite interactions, and, in recent decades, genomics. These studies and genomic resources for them are widely distributed and often widespread in various databases. In this chapter, we analyze the state of the art for genomic resources for Lepidoptera in GenBank for the following genes: *elongation factor-1α*, *wingless*, *cytochrome c oxidase I*, *ribosomal DNA and RNA*, and in general a number of other protein and enzyme entries; complete mitochondrial genomes; complete nuclear genomes; and published work on barcode methodology. This information will help researchers find gaps in the available resources and direct research efforts in these areas.

Abbreviations

cDNA	Complementary DNA
BAC	Bacterial artificial chromosome
CDS	Coding sequences
COI, COII, COIII	*Cytochrome oxidase subunits I, II, III*
cyt b	*Cytochrome b*
dsx	*Doublesex*
EF	*Elongation factor-1α*
EST	Expressed sequence tag

C. Pozo • B. Prado
Departamento de Conservación de la Biodiversidad, El Colegio de la Frontera Sur, Chetumal, Quintana Roo, México
e-mail: cpozo@ecosur.mx; brp_c@yahoo.com

A.N. Castañeda-Sortibrán (✉)
Departamento de Biología Celular, Facultad de Ciencias, Universidad Nacional Autónoma de México, Ciudad de México, México
e-mail: nitxin@ciencias.unam.mx

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*, Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_3

mtDNA	Mitochondrial DNA
MT-ND4L	Mitochondrially encoded NADH dehydrogenase 4L
MT-ND1	Mitochondrially encoded NADH dehydrogenase subunit 1
ncDNA-*18S rRNA*	Nuclear DNA of the small subunit ribosomal RNA
ncDNA-*28S rRNA*	Nuclear DNA of the large subunit ribosomal RNA
NCBI	National Center for Biotechnology Information
rDNA	*Ribosomal DNA*
rRNA	*Ribosomal RNA*
tRNA-Leu	tRNA-leucine
tRNA-Val	tRNA-valine
Wg	*Wingless*
WGS	Whole-genome shotgun

3.1 Introduction: Why Butterflies and Moths?

Lepidoptera is one of the largest groups of organisms in the world. This order comprises insects commonly known as butterflies and moths. Historically, the former have attracted the attention of professional and amateur entomologists, as well as the general public because of the beautiful colors and patterns present in their scaled wings. The moths are studied primarily not only because many species are economically important pests of agriculture and forestry but also for silk production, with the mulberry silkworm, *Bombyx mori*, considered one of the few "domesticated" insects [1], reared at least since 2600 BC [2].

The origin of the holometabolous order Lepidoptera is dated to the Late Carboniferous, but diversification occurred in the Early Cretaceous at the same time as the radiation of flowering plants [3]. Currently, the order Lepidoptera contains over 157,424 species including approximately 22 fossils; the living species (157,402) are classified into 45 superfamilies, 134 families, and 15,562 genera [4]. This is the second most diverse group of animals after Coleoptera.

Insects have long been used as model systems, and the fruit fly, *Drosophila melanogaster*, was the first choice historically, primarily because of its short life cycle and ease of rearing in the laboratory [5]. Nevertheless, the importance of model systems is that discoveries and implications can be extended far beyond the particular organism under study [6]. Certain phenomena such as evolution, coevolution, and biogeographic and ecological mechanisms are better documented and explained within Lepidoptera because there is significant background in the knowledge of this group, mostly due to its economic importance and attractiveness. This gives an advantage to Lepidoptera, as they are better known in many aspects than other diverse groups, and their genomic research will help to understand different kinds of processes.

3.2 GenBank Database: Lepidoptera Representation

In 1982, GenBank was officially released; by 1992, the National Center for Biotechnology Information (NCBI), which is part of the International Nucleotide Sequence Database Collaboration (INSDC), took responsibility for it. From August 2011 to 2012, GenBank had an annual increase in records of 33.1 %, but invertebrates had a decrease of 1.7 % in the same year. The GenBank Dataset is divided into two main groups, taxonomic and functional. The functional division in GenBank sequences makes the data easy to handle and reflects the methods used to obtain it [7]. Functional divisions in 2012 included transcriptome shotgun data, whole-genome shotgun (WGS) data, patented sequences, genome survey sequences, expressed sequence tags (ESTs), high-throughput genomics, sequence tagged sites, and high-throughput complementary DNA (cDNA). Transcriptome shotgun data was the fastest growing division, with more than 200 % growth that year [7]. The taxonomic division, GenBank Dataset, was useful only to know the species of Lepidoptera reported in GenBank. A search in GenBank with “Lepidoptera” on April 2, 2014, returned 1,093,006 sequences; 57,906 registers of these were not identified, yielding 1,035,100 sequences representing a comprehensive landscape of Lepidoptera genomics. According to a recent classification of Lepidoptera [4], 92 % of the 134 living families are represented in GenBank with at least one sequence (Fig. 3.1a), and only 10 families are not present (Anomosetidae, Schistonoeidae, Syringopaidae, Coelopoetidae, Epimarptidae, Whalleyanidae, Simaethistidae, Ratardidae, Peleopodidae, and Metarbelidae). As we go to lower taxonomic categories, the representation in GenBank is reduced to 41 % at the genus level (Fig. 3.1b) and only 13 % at the species level (Fig. 3.1c). Additionally, there is a dissimilarity in the proportion of representation of genera and species from different families or what Wilson [8] observed as uneven taxonomic distribution. Almost 20 % of the families have all their genera represented in GenBank (26 families with 100 %, Fig. 3.2a), including two butterfly families and the rest moths (Table 3.1). Nearly 20 % of the families have less than 20 % of their genera represented. At the species level, representation is very low, with just two families, Carthaeidae and Prodidactidae (Table 3.2), with 100 % representation for only one species each (*Carthaea saturnioides* and *Prodidactis mystica*, respectively), and 65 % of the families with less than 10 % of the species represented (Fig. 3.2b). The family Prodoxidae has proper representation with nearly 80 % of the species, and seven families are 56.8 % represented (Sphingidae, Aididae, Papilionidae, Agathiphagidae, Heterogynidae, Lophocoronidae, and Millieriidae) (Table 3.2).

In total, 124 families, 6336 genera, and 20,076 species of Lepidoptera are represented in GenBank; but a key question is, what functional sequences are documented for each one? We will present information on this using some well-represented sequences for Lepidoptera as a whole.

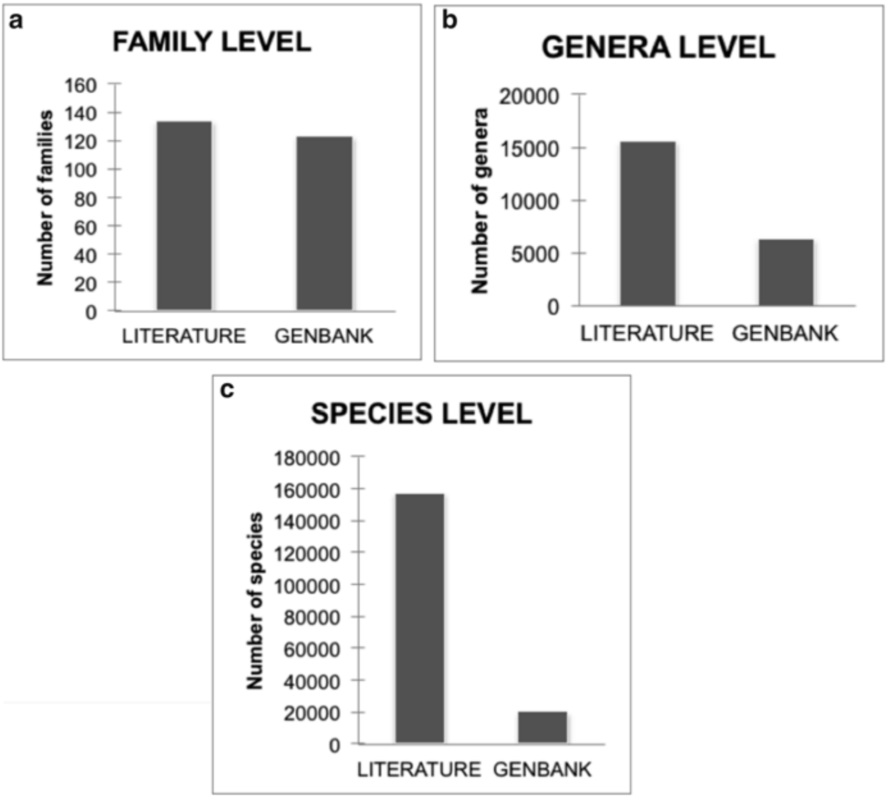


Fig. 3.1 Records of Lepidoptera in GenBank by taxonomic level. (**a**) Comparison between the number of families of Lepidoptera reported by Nieukerken et al. [4] and the families in GenBank as of April 2014. The data represent 92 % of the families of the order. (**b**) Representation at the genus level: only 41 % of the group is represented. (**c**) Representation at the species level: only 13 % of all species of Lepidoptera are represented in GenBank

3.3 Global Lepidoptera Sequences

There are several uses for DNA sequences, such as phylogenetic studies, pest control applications, and analysis of evolutionary changes at the species level and even in particular gene families. Targets of analysis depend on the aims of the research. For instance, different regions of mitochondrial DNA such as *cytochrome oxidase subunits I*, *II*, and *III* (*COI*, *COII*, *COIII*), *cytochrome b* (*cyt b*), or nuclear DNA sequences, e.g., *ribosomal RNA* (*rRNA*), *ribosomal DNA* (*rDNA*), satellite DNA, introns, and nuclear protein-coding genes, can be used to delimit species, phylogeny, or functional genetics [9].

Knowing the nature of the DNA can provide new insights into the biology of this order. The most represented Lepidoptera genes in GenBank are *elongation factor-1α* (*EF*), *wingless* (*Wg*), *rRNA*, *rDNA*, *COI*, and selected proteins. In this chapter,

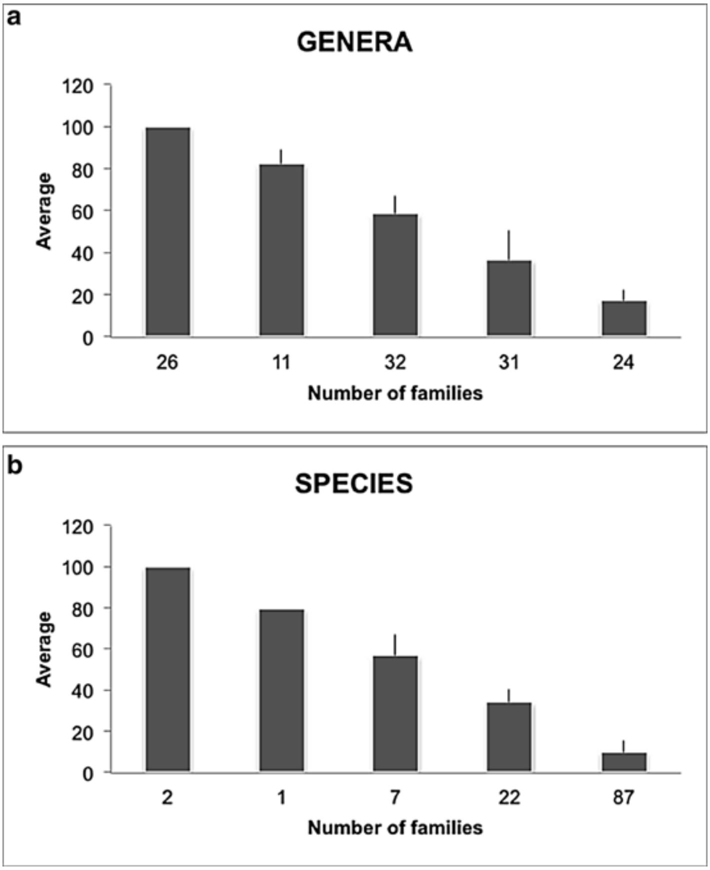


Fig. 3.2 Records of lepidopteran families, genera, and species in GenBank sequence accessions. (**a**) Number of families and average number of genera found in GenBank as of April 2014. There are 26 families containing 100 % of genera. (**b**) At the species level, 87 families have only 10 % of their total species

proteins with a catalytic function are classified as enzymes and the rest remain as proteins.

3.4 *Elongation Factor-1α*

EF is a slowly evolving nuclear gene which is involved in the production of proteins, operating at the receptor site of the ribosome during the translation process [10]. In insects, when used in combination with mitochondrial genes [11, 12], it results in good resolution of high-level phylogenetic relationships, particularly in Lepidoptera [13–19]. Wahlberg et al. [20] resolved the polyphyletic nature of

Table 3.1 Percentage of genera by family in GenBank

	Lepidoptera genera in GenBank by family (percentage)				
	100	82.4	58.7	36.7	17.5
Butterflies	Hedylidae	Nymphalidae	Hesperiidae	Riodinidae	
	Papilionidae	Pieridae	Lycaenidae		
Moths	Adelidae	Callidulidae	Acanthopteroctetidae	Blastobasidae	Alucitidae
	Agathiphagidae	Cecidosidae	Apatelodidae	Brachodidae	Autostichidae
	Aididae	Choreutidae	Bombycidae	Carposinidae	Batrachedridae
	Andesianidae	Coleophoridae	Brahmaeidae	Castniidae	Cosmopterigidae
	Anthelidae	Oecophoridae	Bucculatricidae	Crambidae	Cossidae
	Argyresthiidae	Oenosandridae	Cimeliidae	Dalceridae	Dudgeoneidae
	Attevidae	Phiditiidae	Copromorphidae	Drepanidae	Endromidae
	Bedelliidae	Saturniidae	Doidae	Elachistidae	Epicopeiidae
	Carthaeidae	Xyloryctidae	Eriocottidae	Epermeniidae	Heliocosmidae
	Chimabachidae		Euteliidae	Epipyropidae	Lasiocampidae
	Cyclotornidae		Gracillariidae	Erebidae	Lecithoceridae
	Douglasiidae		Hepialidae	Eupterotidae	Limacodidae
	Eriocraniidae		Hyblaeidae	Gelechiidae	Lyonetiidae
	Galacticidae		Immidae	Geometridae	Momphidae
	Heterobathmiidae		Lacturidae	Glyphipterigidae	Notodontidae
	Heterogynidae		Lypusidae	Heliodinidae	Psychidae

	Himantopteridae		Micropterigidae	Heliozelidae	Pyralidae
	Lophocoronidae		Millieriidae	Incurvariidae	Scythrididae
	Mnesarchaeidae		Mimallonidae	Megalopygidae	Sesiidae
	Prodidactidae		Neopseustidae	Nolidae	Stathmopodidae
	Prodoxidae		Nepticulidae	Opostegidae	Tineidae
	Prototheoridae		Noctuidae	Palaeosetidae	Tortricidae
	Sphingidae		Palaephatidae	Plutellidae	Yponomeutidae
	Tischeriidae		Phaudidae	Pterophoridae	Zygaenidae
			Praydidae	Roeslerstammiidae	
			Pterolonchidae	Somabrachyidae	
			Schreckensteiniidae	Thyrididae	
			Sematuridae	Uraniidae	
			Tineodidae	Ypsolophidae	
			Urodidae		

Table 3.2 Percentage of species by family in GenBank

	Species in GenBank by family (percentage)							
	100	79.59	56.85	34.07	10			
Butterflies			Papilionidae	Hedylidae	Hesperiidae			
				Nymphalidae	Lycaenidae			
				Pieridae	Riodinidae			
Moths	Carthaeidae	Prodoxidae	Agathiphagidae	Acanthopteroctetidae	Adelidae	Dalceridae	Himantopteridae	Phiditiidae
	Prodidactidae		Aididae	Andesianidae	Alucitidae	Douglasiidae	Hyblaeidae	Plutellidae
			Heterogynidae	Anthelidae	Apatelodidae	Drepanidae	Immidae	Praydidae
			Lophocoronidae	Castniidae	Argyresthiidae	Dudgeoneidae	Incurvariidae	Psychidae
			Millieriidae	Cecidosidae	Attevidae	Elachistidae	Lacturidae	Pterophoridae
			Sphingidae	Chimabachidae	Autostichidae	Endromidae	Lasiocampidae	Pyralidae
				Cimeliidae	Batrachedridae	Epermeniidae	Lecithoceridae	Roeslerstammiidae
				Cyclotornidae	Bedelliidae	Epicopeiidae	Limacodidae	Schreckensteiniidae
				Eriocraniidae	Blastobasidae	Epipyropidae	Lyonetiidae	Scythrididae
				Galacticidae	Bombycidae	Erebidae	Lypusidae	Sematuridae
				Heterobathmiidae	Brachodidae	Eriocottidae	Megalopygidae	Sesiidae
				Micropterigidae	Brahmaeidae	Eupterotidae	Mimallonidae	Somabrachyidae
				Mnesarchaeidae	Bucculatricidae	Euteliidae	Momphidae	Stathmopodidae
				Oecophoridae	Callidulidae	Gelechiidae	Neopseustidae	Thyrididae
				Oenosandridae	Carposinidae	Geometridae	Nepticulidae	Tineidae
				Saturniidae	Choreutidae	Glyphipterigidae	Noctuidae	Tischeriidae
				Tineodidae	Coleophoridae	Gracillariidae	Nolidae	Tortricidae
				Xyloryctidae	Copromorphidae	Heliocosmidae	Notodontidae	Uraniidae
				Yponomeutidae	Cosmopterigidae	Heliodinidae	Opostegidae	Urodidae
					Cossidae	Heliozelidae	Palaephatidae	Ypsolophidae
					Crambidae	Hepialidae	Phaudidae	Zygaenidae

Limenitidinae in a cladistic analysis using one mitochondrial gene sequence (*COI*, 1450 bp) and two nuclear gene sequences (*EF*, 1064 bp and *Wg*, 412–415 bp).

The order Lepidoptera has 10,045 sequences of *EF* in GenBank; the Nymphalidae family is the most represented with 2982 entries, followed by Lycaenidae (850), Geometridae (704), Noctuidae (675), Gracillariidae (573), Prodoxidae (485), Erebidae (449), Papilionidae (397), Sphingidae (310), Nepticulidae (298), Pieridae (268), Cosmopterigidae (247), Hesperiidae (234), Tortricidae (173), Crambidae (156), Nolidae (141), and Saturniidae (120) (Fig. 3.3a, Table 3.3). Nymphalidae occupies the first place in the number of genera and species (450 and 1555, respectively). In the second place, Geometridae has only 25 % of the Nymphalidae species, with 390 species in 215 genera (Fig. 3.3a). Butterfly families Papilionidae, Pieridae, and Nymphalidae have a high percentage of genera with *EF* in GenBank, with 90 %, 85 %, and 80 %, respectively.

3.5 *Wingless*

Wg is a nuclear protein-coding gene involved in wing, gut, and nervous system development in insects. In Lepidoptera, it handles the color and spotted pattern of the wing and thus has a critical role in ecological and evolutionary processes [21–24]. It was thought that *Wg* contributed to mimicry, but Kunte et al. [25] recently showed that *Doublesex* (*dsx*) is a mimicry "supergene" involved in female-specific mimicry in *Heliconius* and *Papilio* spp.

Wg has been used to resolve species and subfamily relationships in Nymphalidae [26] and was useful at a tribe level in Riodinidae and Lycaenidae families [22]. For Hesperiidae, however, the resulting relationships are not congruent with those found using *EF* and *COI* [27]. In the Geometridae family, the use of *Wg* in combination with *EF* and three other nuclear genes helped to elucidate the evolution of female flightlessness in the tribe Operophterini [28].

GenBank has 6272 records of lepidopteran *Wg* sequences; Nymphalidae has approximately 40 % of the records, followed by Lycaenidae, Hesperiidae, Erebidae, and Pieridae, with just 5 %. The best-known families based on the number of genera and/or species with records of *Wg* in GenBank are Papilionidae, which have 78.1 % of their genera and 11.2 % of species, and Nymphalidae, with 77 % of their genera and 24 % of species (Fig. 3.3b).

3.6 Enzymes and Proteins

Work with nuclear coding genes such as acetylcholine esterase, alcohol dehydrogenase, actin, chorion, silk genes, and histones, among many others [9], has been significant in Lepidoptera for economic reasons, from silk production in *B. mori* [6, 29, 30] to biological control in pest species like the Asian rice borer, *Chilo*

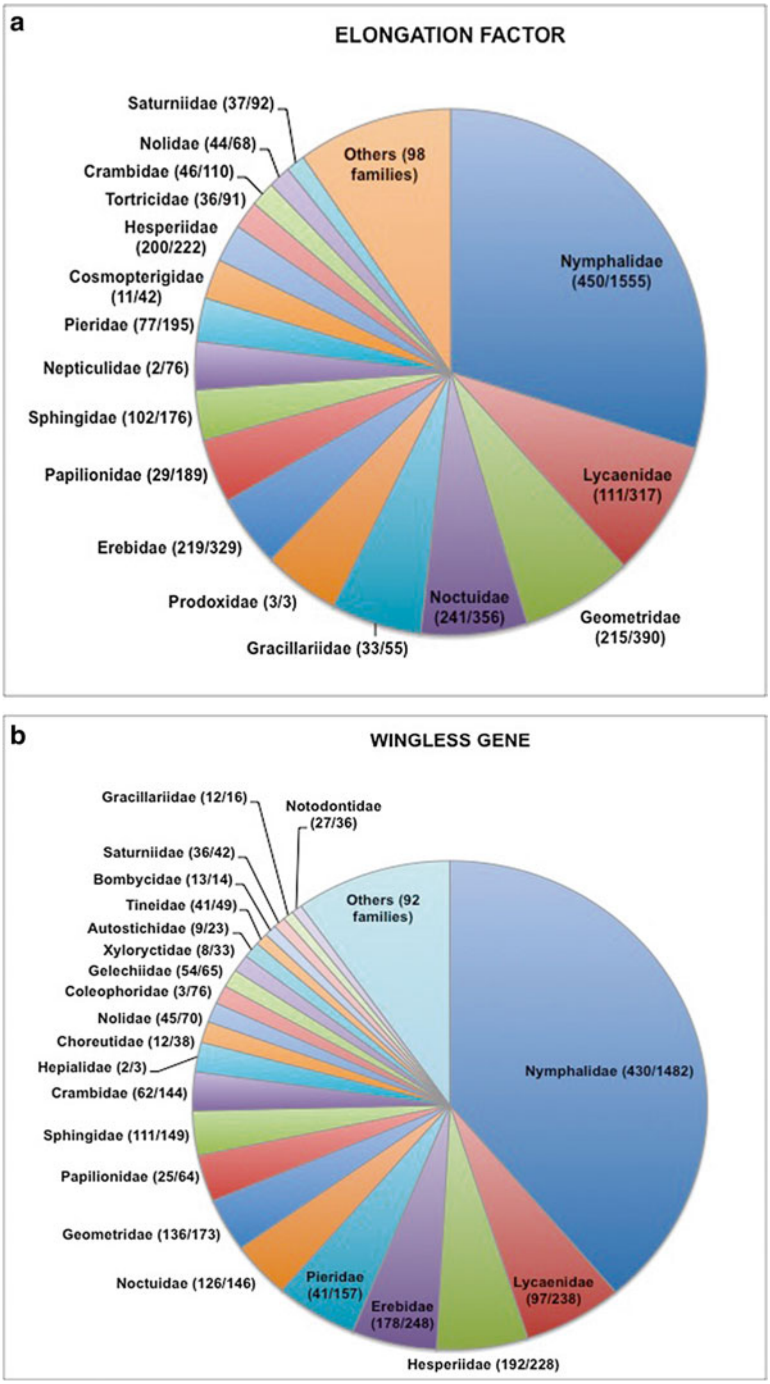


Fig. 3.3 Records of *EF* and *Wg* sequences of Lepidoptera in GenBank. (**a**) Families with *EF* sequenced in GenBank. (**b**) Families with *Wg* sequenced in GenBank. *Numbers* in *brackets* refer to numbers of genera and species

Table 3.3 Summary of families with the most abundant number of sequences in GenBank nuclear and mitochondrial rRNA and rDNA is grouped in "ribosomal"

		Number of records in GenBank						
Family	Genera/species[a]	*COI*	Complete mitogenome	Enzymes	Proteins	*EF*	*Wg*	Ribosomal
Nymphalidae	559/6152	17821	127	8053	26852	2982	2437	2891
Noctuidae	1089/11772	20083	19	2581	2064	675	236	922
Crambidae	1020/9655	13734	29	1100	1072	156	161	396
Bombycidae	26/185	1129	52	5581	26515	58	48	486
Papilionidae	32/570	2858	20	1855	5387	397	193	762
Hesperiidae	570/4113	16091	8	213	88	234	359	155
Erebidae	1760/24569	23939	10	871	170	449	335	496
Sphingidae	206/1463	7375	4	558	538	310	184	120
Geometridae	2002/23002	28329	4	875	452	704	221	730
Notodontidae	704/3800	11662	4	187	53	37	42	19
Tortricidae	1701/10387	8242	19	647	512	173	40	293
Lycaenidae	416/5201	6354	14	792	302	850	392	1143
Gracillariidae	101/1866	2360	0	1276	203	573	42	269
Pieridae	91/1164	3602	17	1078	475	268	324	662

[a]Nieukerken et al. [4]

suppressalis [31], and the tobacco hornworm, *Manduca sexta* [32–34], another important lepidopteran model for basic research (see below). It has also been very important in the study of metabolism associated with life history traits such as diapause and eclosion, as well as the study of metabolic pathways and the structure of proteins [6]. However, even more importantly, protein-coding genes are essential for the resolution of deep phylogenetic branches in Lepidoptera [35–37] and study of evolution in families of genes or domestication events, as in the *Bombyx* genus [38].

GenBank contains 33,268 enzyme sequences for Lepidoptera; the family Nymphalidae is the most represented with 8053 sequences in 382 genera and 978 species, followed by Bombycidae (5581 sequences, 14 genera, and 18 species), Noctuidae (2581 sequences, 236 genera, and 329 species), Papilionidae (1855 sequences, 39 genera, and 225 species), Gracillariidae (1276 sequences, 48 genera, and 77 species), Crambidae (1100 sequences, 338 genera, and 799 species), and Pieridae (1078 sequences, 23 genera, and 85 species) (Fig. 3.4a).

Proteins other than enzymes are documented in GenBank with twice the number of enzyme sequences (67,334 sequences); again, the most represented is Nymphalidae, with 26,852 sequences corresponding to 73 genera and 215 species, followed by Bombycidae (26,515 sequences, 15 genera, and 19 species), Papilionidae (5387 sequences, 5 genera, and 21 species), Noctuidae (2064 sequences, 32 genera, and 53 species), and Crambidae (1072 sequences, 32 genera, and 40 species) (Fig. 3.4b).

3.7 Ribosomal DNA and RNA

Ribosomes are involved in protein synthesis. Eukaryotes contain two major cytoplasmic rRNA subunits, 28S and 18S; tandem arrays of rDNA genes encoding both subunits are located on the nuclear chromosomes, but there are also rDNA genes in the mitochondria (16S and 12S). Genes encoding rRNA have been widely used in phylogenetic analysis because their different regions have distinct rates of evolution, giving diverse resolution for phylogenetic inference [9, 39]. In Lepidoptera, diverse phylogenetic analyses have included mitochondrial *rDNA* to construct phylogeny [40–42]. The term ribosomal is used here to report either mitochondrial or nuclear *rRNA* and *rDNA* sequences,.

GenBank has 11,652 ribosomal accessions, but these include less than 5 % of the total sequences for Lepidoptera. Nymphalidae has the highest numbers of ribosomal records in GenBank (2891), followed by Lycaenidae (1143), Noctuidae (922), and Zygaenidae (895) (Fig. 3.5a). Additionally, Nymphalidae has the highest number of genera and species represented (383 and 1129, respectively), and Papilionidae has 90 % of their genera and 33 % of species represented in GenBank, followed by Nymphalidae (68.5 % genera and 18 % species). Being a small family, it is interesting that Zygenidae appears in the 4th place for the number of accessions for ribosomal sequences in GenBank, where it is represented by 18 genera and 108 species with 895 records. One genus, *Zygaena*, comprises 826 records of ribosomal

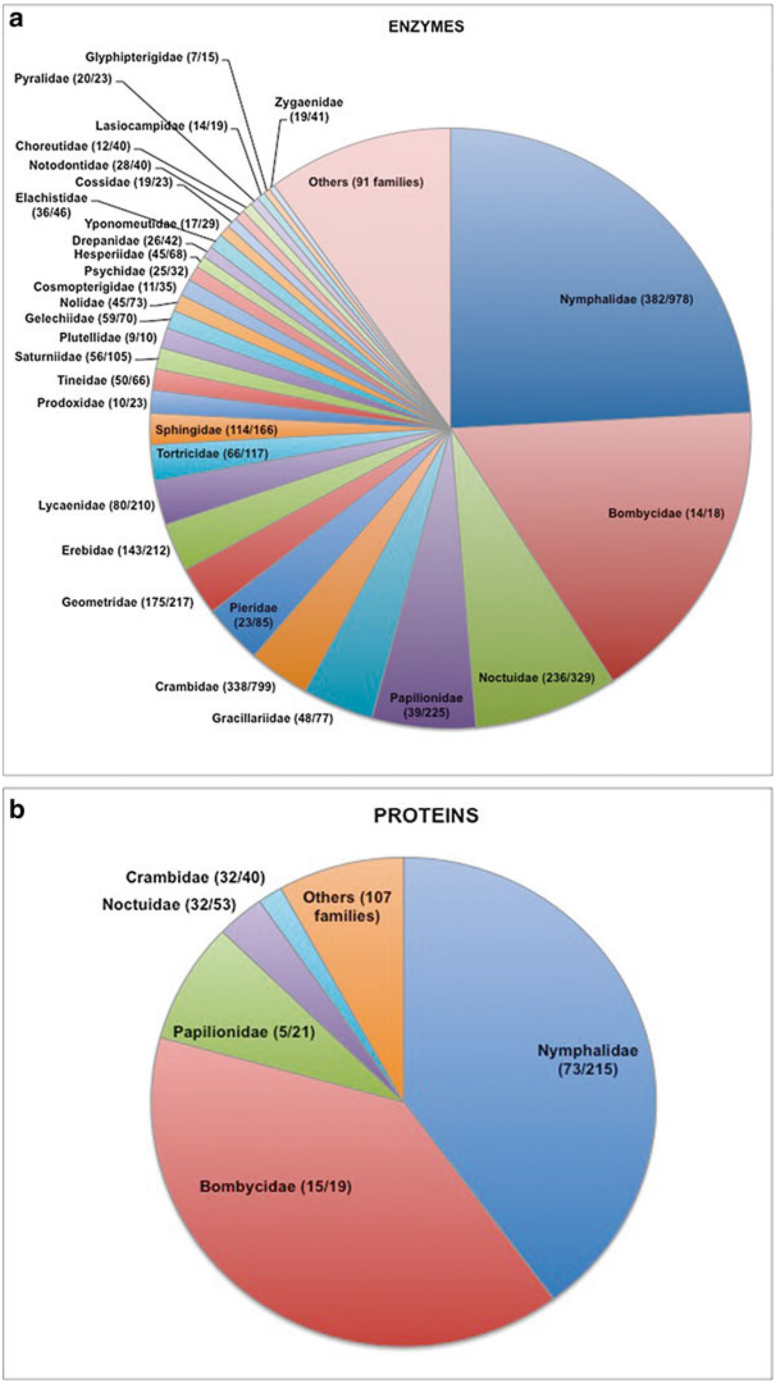


Fig. 3.4 Records of enzyme sequences of Lepidoptera in GenBank. (**a**) Records of Lepidoptera by family that have sequenced enzymes in GenBank and (**b**) families with sequenced proteins in GenBank. The first number between *brackets* refers to the number of genera, and the second is the number of species

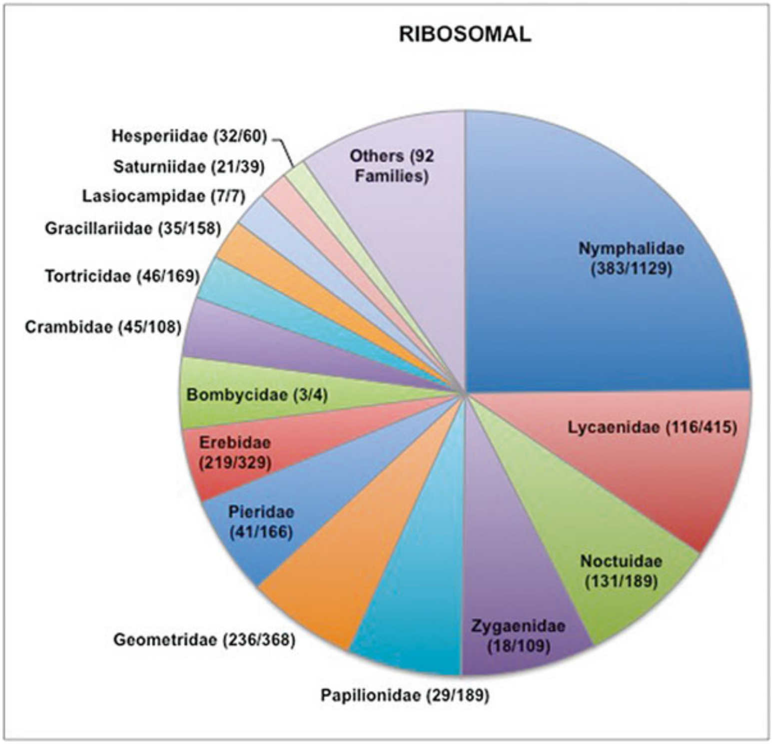


Fig. 3.5 Records of sequenced ribosomal (nuclear and mitochondrial rDNA and rRNA) genes of Lepidoptera in GenBank by family. The first number between *brackets* refers to the number of genera, and the second is the number of species

sequences for 85 species, including 344 records for *Zygaena transalpine* and 125 records for *Z. angelicae* [43]; Niehuis et al. [44] contributed, with the complete sequences of mitochondrially encoded NADH dehydrogenase subunit 1 (*MT-ND1)*, tRNA-leucine (*tRNA-Leu*), 16S rRNA, tRNA-valine (*tRNA-Val*), and, with large fragment of 12S rRNA, nuclear DNA of the small and large subunits ribosomal RNA (ncDNA-*18S rRNA* and ncDNA-*28S rRNA*) for a phylogenetic study of the zygaenoid group.

3.8 *Cytochrome C Oxidase Subunit I (COI)*

Cytochrome c oxidase is a protein complex (subunits 1–3) located in the mitochondria that plays an important role as a terminal enzyme in the respiratory chain, transferring electrons and reducing oxygen to water. This process is carried out by subunit 1 (*COI*) of the complex [45, 46]. Genes encoding *COI* form part of the mitogenome, and analysis of its complete sequence shows that different regions

evolve at distinct rates, making *COI* very useful for insect phylogenetic studies [47]. In Lepidoptera, *COI* by itself has a better resolution at lower levels, such as species and species groups [48]. At higher levels, it is recommended to use *COI* together with other gene sequences (e.g., *Wg*, *EF*) for phylogenetic analysis and dating of divergence times [20, 42, 49, 50]. Given that *COI* has low intraspecific variability and high interspecific variability, it is suitable for species recognition, and in 2003, it was proposed to be used for a universal barcoding system in species identification [51, 52]. The critical sequence consists of an approximately 600 bp long fragment of *COI* which is amplified by PCR and sequenced. Then, this sequence is compared to a library of *COI* sequences of species identified previously by taxonomists. The advantages of using *COI* as a barcoding system include the large number of DNA copies per cell, its maternal inheritance, and lack of introns. In Lepidoptera, the barcoding system works very well, especially for the discovery of new species in groups with crypticism [53–57] and overlooked species [58]. Since the barcoding proposal in 2003, *COI* sequences have been increasing, and as of April 2014, GenBank had 215,074 accessions, which represent 22 % of all the sequences within families of Lepidoptera.

Wilson [8] used a fragment of *COI* (DNA barcode) and two other gene regions (*EF* and *Wg*) of 977 species from Lepidoptera to probe phylogenetic signal and concluded that the DNA barcode fragment has low signal for levels above genus. In the first quarter of 2014, there were 19,279 named species belonging to 6147 genera for *COI* alone; the huge increase in the number of species found in GenBank represents the widespread use of this marker in taxonomic and phylogenetic studies. In fact, GenBank contains 92 % of the lepidopteran families reported by Nieukerken et al. in 2011 [4] and 39.5 % of the genera, but just 12.25 % of the number of species. The Geometridae family has the largest number of genera represented by this gene, followed by Erebidae, Noctuidae, and Nymphalidae. Although Geometridae has the highest number of species, Nymphalidae has more species represented than Erebidae or Noctuidae (Fig. 3.6a). Considering the number of genera reported for each of the families with relatively high numbers of sequences registered in GenBank, coverage of Sphingidae is 99.5 %, followed by Papilionidae, Nymphalidae, Pieridae, and Noctuidae (94 %, 86 %, 77 %, and 67 %, respectively). This pattern is similar at the species level, but Erebidae, with the largest species number reported [4], has only 8.5 % representation in GenBank (Table 3.3 and Fig. 3.6a).

3.8.1 COI *and Barcode Publications in ISI Web of Science and Scopus*

In the period from 2003 to 2013, the total number of publications returned in the ISI Web of Science and Scopus based on a search using keywords "barcode/barcoding Lepidoptera" was 352. The year with the largest number of publications is 2012 (56 papers). The number of publications using barcodes appears to cycle, the first being bigger than the second, with a tendency to increase from 2003 to 2008 with 47

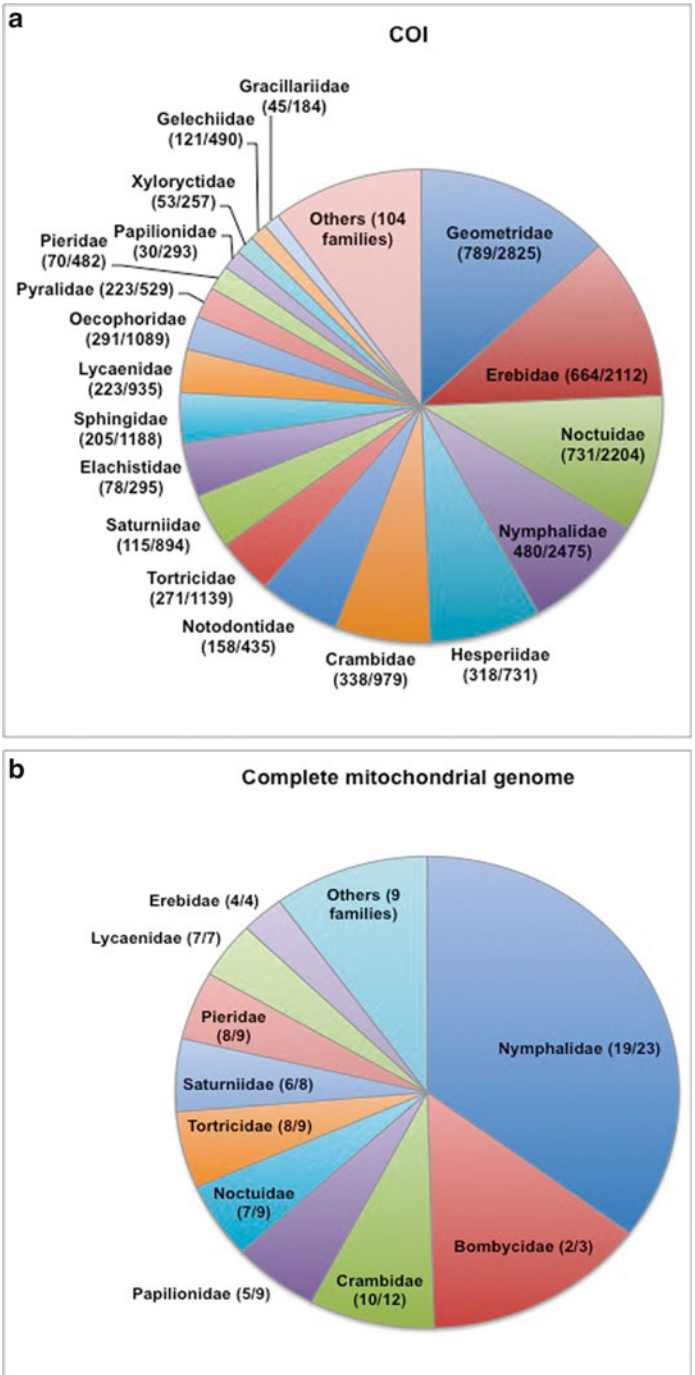
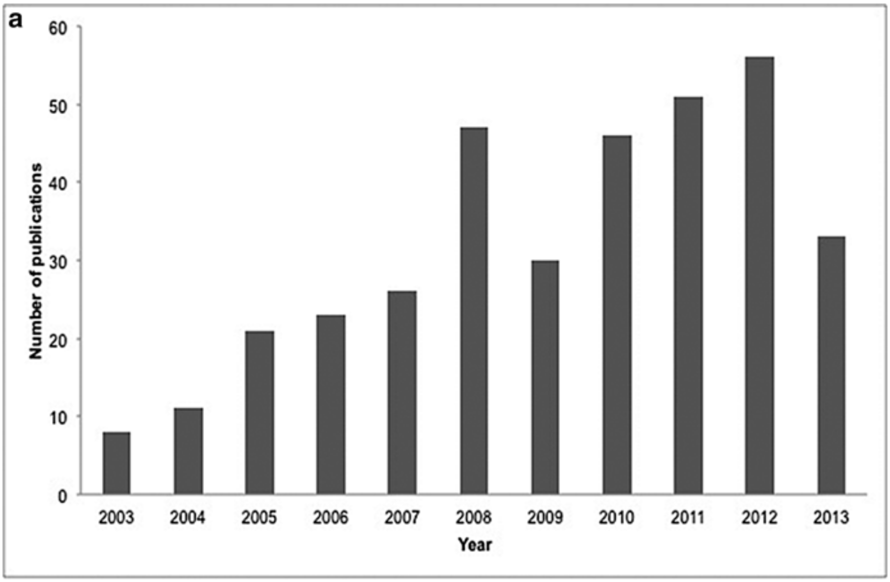


Fig. 3.6 Records of lepidopteran mitochondrial sequences in GenBank. (**a**) *COI* records in GenBank by family of Lepidoptera as of April 2014. (**b**) Families that have a complete mitochondrial genome in GenBank. The first number in *brackets* refers to the number of genera, and the last is the number of species in each family

publications. The second cycle starts in 2009 with a reduction of 36 % and reaches the maximum in 2012 (Fig. 3.7a). These fluctuations are explained by the discovery of new species with crypticism using barcoding and the large inventories of newly detected species, all waiting for a taxonomist to name them in a publication. The type of journal confirms the latter hypothesis, with the largest number of articles on the subject published in *Zootaxa* (28), followed by *Molecular Phylogenetics and Evolution* (24) and *Annals of the Entomological Society of America* (20) (Fig. 3.7b).



b

JOURNAL	NUMBER OF PUBLICATIONS
Zootaxa	28
Molecular Phylogenetics & Evolution	24
PlosOne	20
Annals of the Entomological Society of America	18
European Journal of Entomology	13
Molecular Ecology	12
Molecular Ecology Resources	10
Journal of Economic Entomology	9
Molecular Biology Reports	9
Systematic Entomology	9
Zookeys	8
Journal of Biogeography	7
Zoological Journal of the Linnean Society	7
Florida Entomologist	6
Neotropical Entomology	5
Proceedings of the National Academy of Sciences of the United States of America	5
Zoologica Scripta	5

Fig. 3.7 Publications of lepidopteran *COI* sequences. (**a**) Number of publications of Lepidoptera using *COI* by year. (**b**) Number of publications of Lepidoptera using *COI* by journal

The scientific publications of this information cover 73 families, with just 60 % of the families with sequences registered in GenBank. Families with the highest number of scientific publications are Nymphalidae (66) and Noctuidae (66). All butterfly families have publications (from Hedylidae with 4 to Nymphalidae with 66), but only 51 % of moth families are present in the barcode literature (66 families, Table 3.4). The Noctuidae family contains the majority of moth barcode publications (66), followed by Tortricidae, Geometridae, Erebidae, and Crambidae (39, 37, 25, and 18 studies, respectively).

The publications with *COI* sequences for barcoding are mainly related to topics in taxonomy, evolution, biogeography, and biodiversity. Considering authors with the highest number of publications, 21 authors have five or more publications in this area (Fig. 3.8). N. Wahlberg currently has the most publications; his main area of research includes the systematics and evolution of the butterfly family Nymphalidae.

Table 3.4 Number of publications with *COI* by family of Lepidoptera returned in ISI Web of Science and Scopus based on a search using keywords "barcode/barcoding Lepidoptera"

	Family	Number of publications
Butterflies	Nymphalidae	66
	Papilionidae	38
	Hesperiidae	27
	Pieridae	27
	Lycaenidae	26
	Riodinidae	5
	Hedylidae	4
Moths	Noctuidae	66
	Tortricidae	39
	Geometridae	37
	Erebidae	25
	Crambidae	18
	Gracillariidae	18
	Pyralidae	18
	Saturniidae	13
	Sphingidae	12
	Gelechiidae	10
	Prodoxidae	10
	Bombycidae	9
	Coleophoridae	9
	Elachistidae	8
	Notodontidae	8
	Oecophoridae	8
	Lasiocampidae	7
	Cosmopterigidae	6
	Drepanidae	6
	Sesiidae	6
	Yponomeutidae	6
	Choreutidae	5
	Tineidae	5

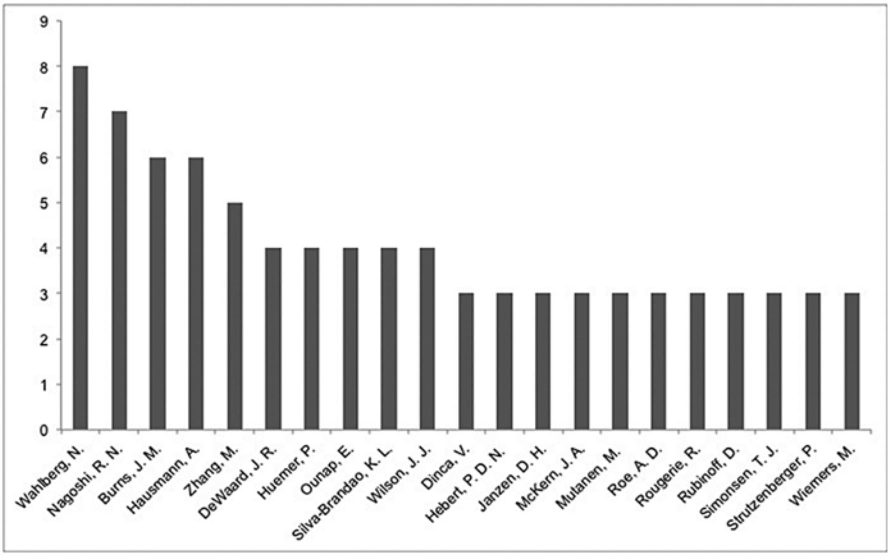


Fig. 3.8 Number of publications of Lepidoptera using *COI* by first authors

3.9 The Complete Mitochondrial Genome

The mitochondrial genome is the most extensively studied genomic system in insects because of its maternal inheritance, lack of recombination, small size, and an accelerated mutation rate compared to nuclear DNA. Mitochondrial DNA (mtDNA) is considerably smaller than nuclear DNA; animal mitochondria are 16–20 kb length, comprising 37 genes and lacking introns [9].

There are distinct regions within mtDNA that diverge at different rates (e.g., *COI*, *COII*, *COIII*, *MT-ND4L* [mitochondrially encoded NADH dehydrogenase 4L], *Cyt b*); as a result, it is very useful at diverse taxonomic levels, even to determine relationships among close species [59]. As noted previously, *COI*, a mitochondrial region of approximately 650 bp, was formally proposed as a barcode system for species identification in 2003 [51, 52]. This and other regions of mtDNA have been used extensively in studies of phylogenetics, comparative and evolutionary genomics, population genetics, molecular evolution, and phylogenomic analysis [60, 61].

Lepidoptera has 361 records of complete mtDNA in GenBank, representing 111 species (as accessed on April 2, 2014). Figure 3.6b shows the proportional representation for families that comprise 90 % of the accessions and the number of genera and species with a mitogenome: Nymphalidae (19/23), Bombycidae (2/3), Crambidae (10/12), Papilionidae (5/9), Noctuidae (7/9), Tortricidae (8/9), Saturniidae (6/8), Pieridae (8/9), Lycaenidae (7/7), and Erebidae (4/4). Nine families represent only 10 % of the accessions. The rapid increase of complete mitochondrial

studies is important; in only 1 month, Wu et al. [62] contributed data for 29 recognized species of Nymphalidae, resulting in a total of 82 species for Papilionoidea and 58 for moths. Now, the largest number of species with complete mitochondrial genomes is Nymphalidae: *Abrota ganga*, *Acraea issoria*, *Apatura ilia*, *A. metis*, *Argynnis childreni*, *A. hyperbius*, *Athyma asura*, *A. cama*, *A. kasa*, *A. opalina*, *A. perius*, *A. selenophora*, *A. sulpitia*, *Bhagadatta austenia*, *Bicyclus anynana*, *Calinaga davidis*, *Danaus plexippus*, *Dichorragia nesimachus*, *Dophia evelina*, *Euploea core*, *E. mulciber*, *Euthalia irrubesens*, *Fabriciana nerippe*, *Heliconius erato*, *H. melpomene*, *H. numata*, *Hipparchia autonoe*, *Issoria lathonia*, *Junonia almanac*, *J. orithya*, *Kallima inachus*, *Libythea celtis*, *Lexias dirtea*, *Melanitis leda*, *M. phedima*, *Melitaea cinxia*, *Neptis philyra*, *N. soma*, *Neope pulaha*, *Pandita sinope*, *Pantoporia hordonia*, *Parantica sita*, *Parasarpa dudu*, *Parthenos sylvia*, *Polyura arja*, *Sasakia charonda*, *S. funebris*, *Sumalia daraxa*, *Tanaecia julii*, *Timelaea maculate*, *Yoma sabina*, and *Ypthima akragas*. The second largest family is Crambidae, a moth family with 12 species: *C. suppressalis*, *Cnaphalocrocis medinalis*, *Diatraea saccharalis*, *Dichocrocis punctiferalis*, *Elophila interruptalis*, *Glyphodes quadrimaculalis*, *Maruca vitrata*, *Ostrinia furnacalis*, *O. nubilalis*, *Paracymoriza distinctalis*, *P. prodigalis*, and *Scirpophaga incertulas*.

Nymphalidae represents the most diverse butterfly family, with 559 genera and 6152 species, which is one-third of all butterfly species [4]. This family has been extensively studied because it includes several species of economic importance as crop pests or potential agents for the biological control of weeds. It is widely distributed in diverse habitats worldwide, and several species have been used as models for ecological, conservation, evolutionary, and developmental studies [63–66]. Nevertheless, the relatively large number of genomic accessions for Nymphalidae is primarily due to many projects related to butterfly phylogeny [62].

Crambidae is a family with some pest species of sod grasses, maize, sugar cane, rice, and other Poaceae, including the sugarcane borer, *D. saccharalis*, which is an economically important pest of several major crops in North and South America. Whole mitogenome sequencing in 2011 was a major step providing molecular markers to monitor changes in population structure associated with acquisition of resistance to *Bacillus thuringiensis*, a class of bacterial endotoxins which is commonly used for pest control [67].

3.10 Genome Projects for Lepidoptera

Knowing the complete genome of Lepidoptera has made it a valuable model system in several ways, including the explanation of key processes such as the immune response, neurophysiology, olfaction, protein biochemistry, evolutionary mechanisms within species (e.g., evolving host–plant utilization) and between species and populations (e.g., wing pattern mimicry), the establishment of phylogenetic relationships, and as a reference for evolutionary comparisons with other insect orders. As of January 2015, eleven lepidopteran genome projects were reported: six

butterflies, of which three are Nymphalidae (*H. melpomene*, *D. plexippus*, and *M. cinxia*) and three Papilionidae (*P. glaucus*, *P. xuthus*, and *P. polytes*), and five moths from diverse families (silk moth, *B. mori* [Bombycidae]; diamondback moth, *Plutella xylostella* [Plutellidae]; rice borer, *C. suppressalis* [Crambidae]; fall army worm, *Spodoptera frugiperda* [Noctuidae]; and tobacco hornworm, *M. sexta* [Sphingidae]) (Table 3.5). The *M. sexta* genome project will be published shortly, along with many other lepidopteran genome projects now in progress (Table 3.6).

Lepidopteran genomes comprise approximately 31 chromosomes [68, 69] with an average size of ~645 Mb, ranging from ~283 Mb (*Danaus plexippus*) to ~1897 Mb (*Euchlaena irraria*) [70]. Sequencing and assembling complete genomes from different lepidopteran species has taken considerable effort compared with the *Drosophila* genome, which has a genome size of ~180 Mb distributed on four chromosomes [71, 72]. Nevertheless, rapid improvements in the actual sequencing techniques and the significance of this group (economical, biological, and ecological) are likely to accelerate sequencing of lepidopteran genomes in order to use them in several ways, such as functional genomics, mutant analysis, bioinformatics, and other post-genomic applications that increase our biological and economical knowledge of Lepidoptera. However, it is important to solve the disaggregation of the community studying Lepidoptera as the great diversity of this group makes it difficult to consolidate operation of a Lepidoptera Consortium, limiting access to major funding [73].

3.10.1 Bombyx mori

The silkworm, *B. mori* (Bombycidae), has been domesticated for silk production for the past 5000 years. It is the most well-studied lepidopteran model system because of its relatively short life cycle [74, 75] and its rich repertoire of well-characterized mutations that affect virtually every aspect of the organism's morphology, development, and behavior. Additionally, it has considerable economic importance. *B. mori* was the first lepidopteran insect genome to be fully sequenced.

In 2004, a Japanese and a Chinese group performed analyses of a WGS draft genome sequence of *B. mori* [76, 77], suggesting that the number of protein-coding genes was 18,000–20,000. The full genome of the silkworm was published in 2008 by the International Silkworm Genome Consortium [78], including a new genome assembly with 16,329 genes. This was made possible by the use of new fosmid- and BAC-end sequence data anchored to a fine genetic map, resulting in an increase in the scaffold size, which made possible a good assembly with low polymorphism (0.2 %) at the nucleotide level.

Based on an extensive database of expressed sequence tags (ESTs) [79] and full-length cDNAs [80], many *Bombyx*-specific genes have been found and annotated, showing the value of transcriptome sequencing for the molecular biology of the silkworm and the whole lepidopteran group.

Table 3.5 Species in GenBank that have a complete genome sequence project, sorted by year of publication

Family	Species	Size (Mb)	Project	Assembly	WGS	Date	Publication
Bombycidae	*Bombyx mori*	481.819	PRJNA205630	GCA_000151625.1	BABH01	23/04/2008	[78]
Nymphalidae	*Danaus plexippus*	272.853	PRJNA72423	GCA_000235995.1	AGBW01	21/11/2011	[82]
Nymphalidae	*Heliconius melpomene*	273.786	PRJNA183487	GCA_000313835.2	CAEZ01	14/02/2012	[87]
Plutellidae	*Plutella xylostella*	393.455	PRJNA78271	GCA_000330985.1	AHIO01	03/01/2013	[91]
Crambidae	*Chilo suppressalis*	314.29	PRJNA178139	GCA_000636095.1	ANCD01	22/04/2014	[95]
Nymphalidae	*Melitaea cinxia*	389.908	PRJNA191594	GCA_000716385.1	APLT01	26/06/2014	[98]
Noctuidae	*Spodoptera frugiperda*	332.567	PRJNA257248	GCA_000753635.2	JQCY02	09/09/2014	[102]
Papilionidae	*Papilio xuthus*	243.89	PRJDB2956	GCA_000836235.1	BBJE01	30/01/2015	[106]
Papilionidae	*Papilio polytes*	227.006	PRJDB2954	GCA_000836215.1	BBJD01	30/01/2015	[106]
Papilionidae	*Papilio glaucus*	374.85	PRJNA270125	GCA_000931545.1	JWHW01	23/02/2015	[105]
Sphingidae	*Manduca sexta*	419.424	PRJNA81037	GCA_000262585.1	AIXA01	13/04/2012	Kanost, et al unpublished

Table 3.6 Species that have a database developed by working groups URLs are provided, although data in some of them could not be updated

Species	Database	URL
Bombyx mori	KAIKObase	http://sgp.dna.affrc.go.jp/KAIKObase/
	Silkworm Genome Database: SilkDB	http://silkworm.genomics.org.cn/
Danaus plexippus	MonarchBase	http://monarchbase.umassmed.edu/
Heliconius melpomene	Heliconius Genome Project	http://butterflygenome.org/
Plutella xylostella	KONAGAbase	http://dbm.dna.affrc.go.jp/px/
Chilo suppressalis	ChiloDB	http://ento.njau.edu.cn/ChiloDB/
Melitaea cinxia	Glanville fritillary butterfly genome project	http://www.helsinki.fi/science/metapop/research/mcgenome.html
Spodoptera frugiperda	SPODOBASE	http://bioweb.ensam.inra.fr/spodobase/
Manduca sexta	Manduca Base	http://agripestbase.org/manduca/
Papilio xuthus and *P. polytes*	PapilioBase	http://papilio.nig.ac.jp/

3.10.2 Danaus plexippus

The monarch butterfly, *D. plexippus* (Nymphalidae), is the most well-recognized species of butterfly, which migrates up to 3000 km from central Mexico to eastern North America [81]. The initial assembly of the monarch genome was made by Zhan et al. in 2011 [82], reporting a genome draft of 273 Mb encoding 16,866 protein-coding genes and suggesting that Lepidoptera is the fastest evolving insect order. In 2013 Zhan et al. [83] established MonarchBase to make the genome data accessible. By 2014, Zhan et al. [84] reported the genetics of monarch butterfly migration and warning coloration, sequencing 80 genomes of *D. plexippus* and nine samples from four additional *Danaus* species. Among other findings, they noted that North American populations are the most basal lineages, with population structure indicating gene flow across North America, and likely origin in the southern USA or northern Mexico. They also found evidence for recurrent, divergent selection on flight muscle function and wing color variation mediated by a myosin gene with no prior known role in insect pigmentation, but an analogous effect in vertebrates. These studies illustrate the power of a genome project to enhance understanding of important biological processes.

3.10.3 Heliconius melpomene

For many years, researchers of the Heliconius group (Nymphalidae) have been searching for the mechanisms underlying adaptive radiation phenomena and Müllerian mimicry. Martin et al. [85] reported interspecific gene flow between sympatric and allopatric populations of *H. melpomene*, *H. cydno*, and *H. timareta*, addressing the idea of evolution without isolation. *H. melpomene* is a model for this type of study, and increased genome research provides the opportunity to explain some of the pathways of adaptive radiation related to the Müllerian mimicry process [86]. The Heliconius Genome Consortium published the *H. melpomene genome* sequence and predicted 12,657 gene models in 2012 [87] and, by comparison with *D. plexippus* and *B. mori*, found the chromosomal organization to be broadly conserved since the Cretaceous. Also, they reported [87] that the genomic region controlling the mimicry pattern has evidence of hybrid exchange of genes between *H. melpomene*, *H. timareta*, and *H. elevatus*. Establishment of this butterfly genome sequence has fuelled significant research, culminating in the recent publication of more robust models for the genetic and mechanistic basis of these phenomena [88].

3.10.4 Plutella xylostella

The diamondback moth, *P. xylostella* (Plutellidae), is one of the more serious pests of cultivated Brassicaceae worldwide [89, 90], which has rapidly evolved high resistance to conventional insecticides such as pyrethroids, organophosphates, fipronil, spinosad, *B. thuringiensis* toxin, and diamides. You et al. [91] published the first whole-genome sequence for this species in 2013, having 18,071 protein-coding and 1412 unique genes with an expansion of gene families related with perception and the detoxification of plant defense chemicals. They found higher levels of *P. xylostella*-specific genes compared with those from *B. mori* (463) and *D. plexippus* (1184). The *P. xylostella*-specific genes are associated with biological pathways essential to monitor and process environmental information, chromosomal replication and/or repair, transcriptional regulation, and carbohydrate and protein metabolism. These authors had to develop special techniques to deal with the extensive polymorphism in the DNA samples because they could not inbreed, as was possible in the other species, or use a cell line, as with *S. frugiperda*. Consequently, the genome was highly fragmented compared to other Lepidoptera genome assemblies. This will be a continuing problem as new sequences are developed for non-model Lepidoptera.

Jouraku et al. [92] developed KONAGAbase, a comprehensive transcriptome database for *P. xylostella*, which can assist researchers in the analysis of genes related to insecticide resistance, allowing the development of more efficient and less environmentally harmful insecticides through clarifying the mechanism of resistance.

3.10.5 Chilo suppressalis

The Asian rice stem borer, *C. suppressalis* (Crambidae), is one of the most economically important pests of rice crops in Northeast China [93]. *C. suppressalis* is a widespread species, extending from East Asia and Oceania into the Middle East and Europe [94]. Given its great economic importance, its metabolism and adaptation to xenobiotics have been extensively studied. In 2014 Yin et al. [95] obtained the first version of a draft genomic sequence for this species using an Illumina sequencing platform to generate WGS sequences that were subsequently assembled. They also established ChiloDB, a database which contains genome and transcriptome sequence data for *C. suppressalis*. In December 2013, they reported the following information was available in ChiloDB: 80,479 scaffolds (length ≥ 2 Kb), 10,221 annotated protein-coding sequences, 262 microRNAs, 82,639 predicted piwi-interacting RNAs, 37,040 midgut transcriptome sequences, 69,977 mixed sample transcriptome sequences, and 77 cytochrome p450 genes or gene fragments. ChiloDB group are working to improve the annotation quality to develop a comprehensive information system for the researchers [95].

3.10.6 Melitaea cinxia

The Glanville fritillary butterfly, *M. cinxia*, belongs to the Nymphalidae family and has been studied to understand the ecological, genetic, and evolutionary consequences of habitat fragmentation on metapopulation dynamics [96]. Vera et al. (2008) [97] reported one of the first studies using 454 pyrosequencing of cDNAs as an approach to genome sequencing for a non-model species and used relatively short sequence assemblies to create a microarray for large-scale functional genomics. However, it was not until 2014 that Ahola et al. [98] sequenced the complete genome of *M. cinxia*, from which they predicted 16,667 gene models. Somervuo et al. (2014) [99] found that a large number of genes were differentially expressed between the landscape types, based on RNA-sequence data. The genome sequence from this lepidopteran, which has the putative ancestral chromosome number (31), provides additional evidence for the evolutionary conservation of lepidopteran chromosomes.

3.10.7 Spodoptera frugiperda

The fall army worm, *S. frugiperda* (Noctuidae), is a polyphagous pest of economic importance in tropical and subtropical countries [100]. Casmuz et al. [101] conducted a literature review of records for this species in North and South America, reporting 186 host plants belonging to 42 different families. This species has devastating effects, damaging crops, and reducing food production [102].

In 2014, the International Centre for Genetic Engineering and Biotechnology (India) used a cell line (Sf9) from the ovary of *S. frugiperda* to obtain a draft sequence of this species. This novel approach gives good results but needs to be validated. Noctuidae is one of the largest families of Lepidoptera containing many of the agriculture pests, and this study represents the first complete genome publication in this family. The genomic DNA was sequenced and assembled into 37,243 scaffolds, 358 Mb in length, with 11,595 predicted genes, of which 36.4 % were assigned a functional characteristic. Repeat elements represent 20.28 % of the total genome. Having the complete genome sequence for this representative of a highly destructive taxonomic group will yield new insights into the evolution of such functions as host–plant specialization, detoxification of allelochemicals, insecticide resistance, and the existence of lepidopteran- and species-specific genes, ultimately helping to understand its biology for improving food production by controlling this species and its close relatives [102].

3.10.8 Papilio glaucus

Species of the genus *Papilio* have been the subject of many evolutionary studies that address issues ranging from population genetics, speciation, and conservation to phylogeny [50]. The North American butterfly, the Eastern tiger swallowtail, *P. glaucus* (Papilionidae), has remarkable morphological and behavioral features that have been described in evolutionary studies, such as Batesian mimicry [103, 104]. High levels of heterozygosity have been a problem in sequencing the genomes of species of Lepidoptera which cannot be easily inbred; the *P. glaucus* genome also has high levels of heterozygosity, similar to *P. xylostella* [105]. Nevertheless, in 2015 Cong et al. [105] succeeded in publishing the complete genome sequence for *P. glaucus* using a single wild-caught individual using a novel assembly strategy. Reporting a genome size of 376 Mb, they predicted 15,695 protein-coding genes and reported the function for 11,975 of them, with repeats constituting 22 % of the genome, values typical of other butterflies.

3.10.9 P. polytes *and* P. xuthus

The common Mormon swallowtail butterfly, *Papilio polytes* (Papilionidae), presents two adult forms, products of a female-limited Batesian mimicry: one mimetic form resembles *Pachliopta aristolochiae* and the other (*cyrus*) is non-mimetic [106]. In 2014, the *dsx* gene was reported by Kunte et al. [25] as a supergene that controls this mimetic expression. This was confirmed in 2015 by Nishikawa et al. [106], who determined whole-genome sequences of *P. polytes* (227 Mb, encoding 12,244 protein-coding genes) and the Asian swallowtail, *P. xuthus* (244 Mb, encoding 13,102 protein-coding genes). Comparison of the sequenced genomes of *P. xuthus*

and *P. polytes* led to the discovery of an extended, highly heterozygous chromosomally inverted region encompassing the genetically mapped locus responsible for the mimetic polymorphism in *P. polyetes* females. The heterozygous, inverted region includes *dsx*, consistent with its proposed involvement in expression of the mimicry pattern. The *Papilio* genome projects are the most recent ones registered in GenBank and the first reports of an association of such a chromosome change with a historically significant phenotype in Lepidoptera. Such a phenomenon is unlikely to have been found without access to the genome sequences.

3.10.10 Manduca sexta

The tobacco hornworm, *M. sexta* (Sphingidae), has been used as a model system for many different fundamental studies of insect and lepidopteran biology, including behavior, immune response, transcription factors, olfaction, biochemistry, physiology, growth, and phylogenetic studies [33, 34, 107–112]. Recently, in 2012, a WGS genome project of *M. sexta* was registered in GenBank by M. Kanost, G. Blissard, J. Qu, S. Richards, et al. (accession number AIXA00000000.1) The genome sequence of this species will lead to an advanced understanding of many basic mechanisms in insect interactions with plants, other insects, and microbes, with potential applications in the areas of biomedicine (insect-vectored diseases) and agriculture (insect–plant interactions). As yet no publications concerning this sequencing project are available but are anticipated in the near future.

3.11 Lepidoptera Genomics Enlightens the Biological Sciences

Butterfly and moth sequences for individual mRNAs were first submitted to GenBank database in the early 1980s [113, 114]. Butterfly and moth genomes, particularly the *B. mori* genome, were among the first insect genomes to be sequenced; the *B. mori* genome was sequenced because of the importance of this insect in silk production, which researchers were focused on improving. Subsequent sequencing of Lepidoptera has targeted other economically significant species, such as *S. frugiperda* and *P. xylostella*. Despite the many GenBank entries (over one million) for the order Lepidoptera, the richness and biological diversity of this order remain underrepresented. The primary aim of current research is to explain complex processes, such as evolution, from a whole-genome perspective, for which lepidopterans are excellent models because many of their ecological and evolutionary traits are known. This potential has already been noticed, and now is the time to use deep genomics to understand these processes. New sequencing technologies are simplifying this task. Further work should focus on obtaining additional species with complete genomes to gain a better representation of the order Lepidoptera in the

GenBank database. Additionally, taxonomists have an important task regarding sequenced specimens that remain unnamed because of the way in which barcoding with *COI* has accelerated the discovery of greater biodiversity. With greater collaborations among ecological, biological, biogeographical, evolutionary, and genomic researchers using Lepidoptera, new findings that will affect fundamental knowledge in all biological sciences can be discovered.

Acknowledgments We thank the reviewers of this chapter, M. R. Goldsmith, A. A. Tolulope, and R. Chandrasekar, who helped us to improve it. In particular, M. R. Goldsmith helped us on the incorporation of information in a relevant way.

We would like to thank Jovana Jasso Martínez, Karen Fernanda Real Salazar, Ana Karina Cruz Galindo, and Azalea Guadalupe Acosta Carreón, who have contributed to this work.

This work was supported by El Colegio de la Frontera Sur and Facultad de Ciencias, Universidad Nacional Autónoma de México.

References

1. Roe AD, Weller SJ, Baixeras J, Brown J, Cummings MP, Davis DR et al (2010) Evolutionary framework for Lepidoptera model systems. In: Goldsmith M, Marec F (eds) Genetics and molecular biology of Lepidoptera. CRC Press, Boca Raton, pp 1–24
2. Scoble MJ (1992) The Lepidoptera: form, function, and diversity. Oxford University Press, New York
3. Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C et al (2014) Phylogenomics resolves the timing and pattern of insect evolution. Science 346:763–767
4. Nieukerken EJV, Kaila L, Kitching IJ, Kristensen NP, Lees DC, Minet J et al (2011) Order Lepidoptera. In: Zhang Z-Q (ed) Animal biodiversity: an introduction to higher-level classification and taxonomic richness. Zootaxa 3148:212–221, Aukland, New Zealand
5. Rubin GM, Lewis EB (2000) A brief history of *Drosophila*'s contributions to genome research. Science 287(5461):2216–2218
6. Willis JH, Wilkins AS, Goldsmith MR (1995) A brief history of Lepidoptera as model systems. In: Goldsmith MR, Wilkins AS (eds) Molecular model systems in the Lepidoptera. Cambridge University Press, Cambridge, pp 1–20
7. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2012) GenBank. Nucleic Acids Res:1–7. doi:10.1093/nar/gks1195
8. Wilson JJ (2010) Assessing the value of DNA barcodes and other priority gene regions for molecular phylogenetics of Lepidoptera. PLoS One 5(5):e10525. doi:10.1371/journal.pone.0010525
9. Hoy MA (2003) Insect molecular genetics. An introduction to principles and applications. Academic, Boston
10. Maroni G (1993) An atlas of *Drosophila* genes. Oxford University Press, Oxford
11. Lin CP, Danforth BN (2004) How do insect nuclear and mitochondrial gene substitution patterns differ? Insights from Bayesian analyses of combined datasets. Mol Phylogenet Evol 30:686–702
12. Kim M, Wan X, Kim MJ, Jeong HC, Ahn N, Kim K et al (2010) Phylogenetic relationships of true butterflies (Lepidoptera: Papilionoidea) inferred from *COI*, *16S rRNA* and *EF-1α* sequences. Mol Cells 30:409–425
13. Cho S, Mitchell A, Regier JC, Mitter C, Poole RW, Friedlander TP et al (1995) A highly conserved nuclear gene for low-level phylogenetics: *Elongation factor-1α* recovers morphology-based tree for heliothine moths. Mol Biol Evol 12:650–656

14. Friedlander TP, Horst KR, Regier JC, Mitter C, Peigler RS, Fang QQ (1998) Two nuclear genes yield concordant relationship within Attacini (Lepidoptera: Saturniidae). Mol Phylogenet Evol 9:131–140
15. Mitchell A, Cho S, Regier JC, Mitter C, Poole RW, Matthews M (1997) Phylogenetic utility of *elongation factor-1 alpha* in noctuidae (Insecta: Lepidoptera): the limits of synonymous substitution. Mol Biol Evol 14(4):381–390
16. Mitchell A, Mitter C, Regier JC (2000) More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analysis of Noctuoidea (Insecta: Lepidoptera). Syst Biol 49:202–224
17. Moulton JK (2000) Molecular sequence data resolves basal divergences within Simuliidae (Diptera). Syst Entomol 25:95–113
18. Regier JC, Mitter C, Peigler RS, Friedlander TP (2000) Phylogenetic relationship in Lasiocampidae (Lepidoptera): initial evidence from *elongation factor-1 alpha* sequences. Insect Syst Evol 31:179–186
19. Caterino MS, Cho S, Sperling FAH (2000) The current state of insect molecular systematics: a thriving Tower of Babel. Annu Rev Entomol 45:1–54
20. Wahlberg N, Weingartner E, Nylin S (2003) Towards a better understanding of the higher systematics of Nymphalidae (Lepidoptera: Papilionoidea). Mol Phylogenet Evol 28:473–484. doi:10.1016/S1055-7903(03)00052-6
21. Carroll SB, Gates J, Keys DN, Paddock SW, Panganiban GE, Selegue JE et al (1994) Pattern formation and eyespot determination in butterfly wings. Science 265(5168):109–114
22. Campbell DL, Brower AV, Pierce NE (2000) Molecular evolution of the *wingless* gene and its implications for the phylogenetic placement of the butterfly family Riodinidae (Lepidoptera: Papilionoidea). Mol Biol Evol 17(5):684–696
23. Beldade P, Brakefield PM (2002) The genetics and evo-devo of butterfly wing patterns. Nat Genet 3:442–452
24. Werner T, Koshikawa S, Williams TM, Carroll SB (2010) Generation of a novel wing colour pattern by the *Wingless* morphogen. Nature 464:1143–1148
25. Kunte K, Zhang W, Tenger-Trolander A, Palmer DH, Martin A, Reed RD et al (2014) *Doublesex* is a mimicry supergene. Nature 507(7491):229–232
26. Brower AVZ, DeSalle R (1998) Patterns of mitochondrial versus nuclear DNA sequence divergence among nymphalid butterflies: the utility of *wingless* as a source of characters for phylogenetic inference. Insect Mol Biol 7:1–10
27. Warren AD, Ogawa JR, Brower AVZ (2008) Phylogenetic relationships of subfamilies and circumscription of tribes in the family Hesperiidae (Lepidoptera: Hesperioidea). Cladistics 24:642–676
28. Snäll N, Tammaru T, Wahlberg N, Viidalepp J, Ruohomaki K, Savontaus ML et al (2007) Phylogenetic relationships of the tribe Operophterini (Lepidoptera, Geometridae): a case study of the evolution of female flightlessness. Biol J Linn Soc 92(2):241–252
29. Fedic R, Zurovec M, Sehnal F (2002) The silk of Lepidoptera. J Insect Biotechnol Sericol 71:1–15
30. Goldsmith MR, Shimada T, Abe H (2004) The genetics and genomics of the silkworm, *Bombyx mori*. Annu Rev Entomol 50:71–100
31. Gong ZJ, Zhou WW, Yu HZ, Mao CG, Zhang CX, Cheng JA et al (2009) Cloning, expression and functional analysis of a general odorant-binding protein 2 gene of the rice striped stem borer, *Chilo suppressalis* (Walker) (Lepidoptera: Pyralidae). Insect Mol Biol 18(3):405–417
32. Feng L, Prestwich GD (1997) Expression and characterization of a lepidopteran general odorant binding protein. Insect Biochem Mol Biol 27(5):405–412
33. Martin JP, Lei H, Riffell JA, Hildebrand JG (2013) Synchronous firing of antennal-lobe projection neurons encodes the behaviorally effective ratio of sex-pheromone components in male *Manduca sexta*. J Comp Physiol A 199:963–979
34. Vogt RG, Große-Wilde E, Zhou J-J (2015) The Lepidoptera odorant binding protein gene family: gene gain and loss within the GOBP/PBP complex of moths and butterflies. Insect Biochem Mol Biol. 62:142–153 http://dx.doi.org/10.1016/j.ibmb.2015.03.003

35. Mutanen M, Wahlberg N, Kalla L (2010) Comprehensive gene and taxon coverage elucidates radiation patterns in moths and butterflies. Proc R Soc B:277(1695):2839–2848. doi:10.1098/rspb.2010.0392
36. Regier JC et al (2009) Toward reconstructing the evolution of advanced moths and butterflies (Lepidoptera: Ditrysia): an initial molecular study. BMC Evol Biol 9:280. doi:10.1186/1471-2148-9-280
37. Regier JC et al (2013) A large-scale, higher-level, molecular phylogenetic study of the insect order Lepidoptera (moths and butterflies). PLoS One 8(3):e58568. doi:10.1371/journal.pone.0058568
38. Xia Q, Guo Y, Zhang Z, Li D, Xuan Z, Li Z et al (2009) Complete resequencing of 40 genomes reveals domestication events and genes in silkworm (*Bombyx*). Science 326(5951):433–436
39. Hills DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66(4):411–453
40. Pashley DP, Ke LD (1992) Sequence evolution in mitochondrial ribosomal and ND-1 genes in lepidoptera: implications for phylogenetic analyses. Mol Biol Evol 9(6):1061–1075
41. Wiegmann BM, Mitter C, Regier JC, Friedlander TP, Wagner DM, Nielsen ES (2000) Nuclear genes resolve Mesozoic-aged divergences in the insect order Lepidoptera. Mol Phylogenet Evol 15(2):242–259
42. Zimmermann M, Wahlberg N, Descimon H (2000) Phylogeny of *Euphydryas* checkerspot butterflies (Lepidoptera: Nymphalidae) based on mitochondrial DNA sequence data. Ann Entomol Soc Am 93(3):347–355
43. von Reumont BJ, Struwe J-F, Schwarzer J, Misof B (2011) Phylogeography of the burnet moth *Zygaena transalpina* complex: molecular and morphometric differentiation suggests glacial refugia in Southern France, Western France and micro-refugia within the Alps. J Zool Syst Evol Res 50(1):38–50. doi:10.1111/j.1439-0469.2011.00637.x
44. Niehuis O, Yen SH, Naumann CM, Misof B (2006) Higher phylogeny of zygaenid moths (Insecta: Lepidoptera) inferred from nuclear and mitochondrial sequence data and the evolution of larval cuticular cavities for chemical defence. Mol Phylogenet Evol 39:812–829
45. Capaldi RA, Malatesta F, Darley-Usmar VM (1983) Structure of cytochrome c oxidase. BBA Bioenergetics 726(2):135–148
46. Michel H (1998) The mechanism of proton pumping by cytochrome c oxidase. Proc Natl Acad Sci U S A 95:12819–12824
47. Lunt DH, Zhang DX, Szymura JM, Hewitt GM (1996) The insect cytochrome oxidase I gene: evolutionary patterns and conserved primers for phylogenetics studies. Insect Mol Biol 5(3):153–165
48. Caterino MS, Sperling FAH (1999) Papilio phylogeny based on mitochondrial cytochrome oxidase I and II genes. Mol Phylogenet Evol 11(1):122–137
49. Brower AVZ (1994) Phylogeny of *Heliconius* butterflies inferred from mitochondrial DNA sequences (Lepidoptera: Nymphalidae). Mol Phylogenet Evol 3(2):159–174
50. Zakharov E, Caterino MS, Sperling FAH (2004) Molecular phylogeny, historical biogeography, and divergence time estimates for swallowtail butterflies of the genus *Papilio* (Lepidoptera: Papilionidae). Syst Biol 53(2):193–215
51. Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. P Roy Soc B Biol Sci 270:313–321
52. Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. P Roy Soc B Biol Sci 270:S596–S599
53. Hebert PDN, Penton EH, Burns JM, Janzen DH, Hallwachs W (2004) Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. Proc Natl Acad Sci U S A 101:14812–14817
54. Janzen DH, Hajibabaei M, Burns J, Hallwachs W, Remigio E, Hebert PDN (2005) Wedding biodiversity inventory of a large and complex Lepidoptera fauna with DNA barcoding. Philos Trans R Soc Lond B Biol Sci 2005 Oct 29; 360(1462):1835–1845

55. Janzen DH, Hallwachs W, Blandin P, Burns JM, Cadiou JM, Chacon I et al (2009) Integration of DNA barcoding into an ongoing inventory of complex tropical biodiversity. Mol Ecol Resour 9:1–26
56. Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2007) DNA barcodes of closely related (but morphologically and ecologically distinct) species of skipper butterflies (Hesperiidae) can differ by only one to three nucleotides. J Lepid Soc 61:138–153
57. Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PDN (2008) DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservacion Guanacaste, Costa Rica. Proc Natl Acad Sci U S A 105:6350–6355
58. Prado B, Pozo C, Valdez-Moreno M, Hebert PDN (2011) Beyond the colours: discovering hidden diversity in the nymphalidae of the Yucatan peninsula in Mexico through DNA barcoding. PLoS One 6(11):e27776. doi:10.1371/journal.pone.0027776
59. Beltran M, Jiggins CD, Bull V, Linares M, Mallet J, McMillan WO et al (2002) Phylogenetic discordance at the species boundary: comparative gene genealogies among rapidly radiating *Heliconius* butterflies. Mol Biol Evol 19(12):2176–2190
60. Hu J, Zhang D, Hao J, Huang D, Cameron S, Zhu C (2010) The complete mithocondrial genome of the yellow coaster, *Craea issoria* (Lepidoptera: Nymphalidae: Heliconiinae: Acraeini): sequence, gene organization and a unique tRNA translocation event. Mol Biol Rep 37:3431–3438
61. Cameron SL (2014) Insect mitocondrial genomics: implications for evolution and phylogeny. Annu Rev Entomol 59:95–117
62. Wu L, Lin L, Lees DC, Hsu Y (2014) Mitogenomic sequences effectively recover relationships within brush-footed butterflies (Lepidptera: Nymphalidae). BMC Genomics 15:468
63. Ackery PR, Vane-Wright RI (1984) Milkweed butterflies: their cladistics and biology, being an account of the natural history of the Danainae, a subfamily of the Lepidoptera, Nymphalidae. British Museum, London
64. Ehrlich PR, Hanski I (2004) On the wings of checkerspots: a model system for population biology. Oxford University Press, New York
65. Sheppard PM, Turner J, Brown K, Benson W, Singer M (1985) Genetics and the evolution of Muellerian mimicry in *Heliconius* butterflies. Philos Trans R Soc Lond B Biol Sci 308:433–610
66. Pollard E, Yates TJ (1993) Monitoring butterflies for ecology and conservation. Chapman and Hall, London
67. Li W, Zhang X, Fan Z, Yue B, Huang F, King E et al (2011) Structural characteristics and phylogenetic analysis of the mitochondrial genome of the sugarcane borer, *Diatraea saccharalis* (Lepidoptera: Crambidae). DNA Cell Biol 30(1):3–8
68. Saura A, von Schoultz B, Saura AO, Brown KS Jr (2013) Chromosome evolution in Neotropical butterflies. Hereditas 150:26–37
69. Robinson R (1971) Lepidoptera genetics. Pergamon, Oxford
70. Gregory TR, Hebert PDN (2003) Genome size variation in lepidopteran insects. Can J Zool 81:1399–1405
71. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG et al (2000) The genome sequence of *Drosophila melanogaster*. Science 287(5461):2185–2195
72. Celniker SE, Rubin GM (2003) The *Drosophila melanogaster* genome. Annu Rev Genom Hum G 4:89–117
73. Beldade P, McMillan WO, Papanicoloau A (2008) Butterfly genomics eclosing. Heredity 100:150–157
74. Goldsmith M, Marec F (2010) Genetics and molecular biology of Lepidoptera. CRC Press, Boca Raton
75. Bisch-Knaden S, Daimon T, Shimada T, Hansson BS, Sachse S (2014) Anatomical and functional analysis of domestication effects on the olfactory system of the silkmoth *Bombyx mori*. P Roy Soc B Biol Sci 281(1774):20132582
76. Mita K, Kasahara M, Sasaki S, Nagayasu Y, Yamada T, Kanamori H et al (2004) The genome sequence of silkworm, *Bombyx mori*. DNA Res 11(1):27–35

77. Xia Q, Zhou Z, Lu C, Cheng D, Dai F, Li B et al (2004) A draft sequence for the genome of the domesticated silkworm (*Bombyx mori*). Science 306(5703):1937–1940
78. The International Silkworm Genome Consortium (2008) The genome of a lepidopteran model insect, the silkworm *Bombyx mori*. Insect Biochem Mol Biol 38(12):1036–1045
79. Mita K, Morimyo M, Okano K, Koike Y, Nohata J, Kawasaki H et al (2003) The construction of an EST database for *Bombyx mori* and its application. Proc Natl Acad Sci U S A 100(24):14121–14126
80. Suetsugu Y, Futahashi R, Kanamori H, Kadono-Okuda K, Sasanuma S, Narukawa J et al (2013) Large scale full-length cDNA sequencing reveals a unique genomic landscape in a lepidopteran model insect, *Bombyx mori*. G3 (Bethesda) 3(9):1481–1492
81. Miller NG, Wassenaar LI, Hobson KA, Norris DR (2012) Migratory connectivity of the monarch butterfly (*Danaus plexippus*): patterns of spring re-colonization in eastern North America. PLoS One 7(3):e31891
82. Zhan S, Merlin C, Boore JL, Reppert SM (2011) The monarch butterfly genome yields insights into long-distance migration. Cell 147(5):1171–1185
83. Zhan S, Reppert SM (2013) MonarchBase: the monarch butterfly genome database. Nucleic Acids Res 41(D1):D758–D763
84. Zhan S, Zhang W, Niitepõld K, Hsu J, Haeger JF, Zalucki MP et al (2014) The genetics of monarch butterfly migration and warning colouration. Nature 514(7522):317–321
85. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F et al (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res 23(11):1817–1828
86. Cuthill JH, Charleston M (2012) Phylogenetic Codivergence supports coevolution of mimetic *Heliconius* butterflies. PLoS One 7(5):e36464. doi:10.1371/journal.pone.0036464
87. Heliconius Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487(7405):94–98
88. Kronforst MR, Papa R (2015) The functional basis of wing patterning in *Heliconius* butterflies: the molecules behind mimicry. Genetics 200:1–19
89. Sarfraz M, Dosdall LM, Keddie BA (2006) Diamondback moth–host plant interactions: implications for pest management. Crop Prot 25(7):625–639
90. De Bortoli SA, Polanczyk RA, Vacari AM, De Bortoli CP, Duarte RT (2013) *Plutella xylostella* (Linnaeus, 1758) (Lepidoptera: Plutellidae): tactics for integrated pest management in Brassicaceae. In: Soloneski S (ed) Weed and pest control–conventional and new challenges. InTech. doi:5772/54110
91. You M, Yue Z, He W, Yang X, Yang G, Xie M et al (2013) A heterozygous moth genome provides insights into herbivory and detoxification. Nat Genet 45(2):220–225
92. Jouraku A, Yamamoto K, Kuwazaki S, Urio M, Suetsugu Y, Narukawa J et al (2013) KONAGAbase: a genomic and transcriptomic database for the diamondback moth, *Plutella xylostella*. BMC Genomics 14(1):464
93. Su JW, Xuan WJ, Sheng CF, Ge F (2003) Biology of overwintering larvae of the Asiatic rice borer, *Chilo suppressalis*, in paddy fields of Northeast China. Entomol Knowl 4:007
94. Khan ZR, Litsinger JA, Barrion AT, Villanueva FFD (1991) World bibliography of Rice Stem Borers 1794–1990. International Rice Research Institute, Makati
95. Yin C, Liu Y, Liu J, Xiao H, Huang S, Lin Y et al (2014) ChiloDB: a genomic and transcriptome database for an important rice insect pest *Chilo suppressalis*. Database:1–7. Published online 2005 Sep 14. doi:10.1098/rstb.2005.1715
96. Hanski I (2011) Eco-evolutionary spatial dynamics in the Glanville fritillary butterfly. Proc Natl Acad Sci U S A 108:14397–14404. doi:10.1073/pnas.1110020108
97. Vera JC, Wheat CW, Fescemyer HW, Frilander MJ, Crawford DL, Hanski I et al (2008) Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. Mol Ecol 17(7):1636–1647
98. Ahola V, Lehtonen R, Somervuo P, Salmela L, Koskinen P, Rastas P et al (2014) The Glanville fritillary genome retains an ancient karyotype and reveals selective chromosomal fusions in Lepidoptera. Nat Commun. 5:4737 doi:10.1038/ncomms5737

99. Somervuo P, Kvist J, Ikonen S, Auvinen P, Paulin L, Koskinen P et al (2014) Transcriptome analysis reveals signature of adaptation to landscape fragmentation. PLoS One 9(7):e101467
100. Valencia Cataño SJ, Rodríguez Chalarca J, Mesa Cobo NC (2014) Effect of varieties of cotton GM on *Spodoptera frugiperda* Smith (Lepidoptera: Noctuidae) larvae. Acta Agron 63:63–70
101. Casmuz A, Juárez ML, Socías MG, Murúa MG, Prieto S, Medina S et al (2010) Revisión de los hospederos del gusano cogollero del maíz, *Spodoptera frugiperda* (Lepidoptera: Noctuidae). Revista de la Sociedad Entomológica Argentina 69:209–231
102. Kakumani PK, Malhotra P, Mukherjee SK, Bhatnagar RK (2014) A draft genome assembly of the army worm, *Spodoptera frugiperda*. Genomics 104(2):134–143
103. Brower JVZ (1958) Experimental studies of mimicry in some North American butterflies: Part II. *Battus philenor* and *Papilio troilus*, *P. polyxenes* and *P. glaucus*. Evolution 12:123–136
104. Clarke CA, Sheppard PM (1962) The genetics of the mimetic butterfly *Papilio glaucus*. Ecology 43:159–161
105. Cong Q, Borek D, Otwinowski Z, Grishin NV (2015) Tiger swallowtail genome reveals mechanisms for speciation and caterpillar chemical defense. Cell Rep 10:910–919. doi:10.1016/j.celrep.2015.01.026
106. Nishikawa H, Iijima T, Kajitani R, Yamaguchi J, Ando T, Suzuki Y et al (2015) A genetic mechanism for female-limited Batesian mimicry in *Papilio* butterfly. Nat Genet 47(4):405–411
107. He Y, Cao X, Li K, Hu Y, Chen YR, Blissard G et al (2015) A genome-wide analysis of antimicrobial effector genes and their transcription patterns in *Manduca sexta*. Insect Biochem Mol Biol 62:23–37. doi:10.1016/j.ibmb.2015.01.015
108. Cao X, He Y, Hu Y, Wang Y, Chen YR, Bryant B et al (2015) The immune signaling pathways of *Manduca sexta*. Insect Biochem Mol Biol 62:64–74. doi:10.1016/j.ibmb.2015.03.006
109. Zhang X, He Y, Cao X, Gunaratna RT, Chen YR, Blissard G et al (2015) Phylogenetic analysis and expression profiling of the pattern recognition receptors: insights into molecular recognition of invading pathogens in *Manduca sexta*. Insect Biochem Mol Biol 62:38–50. doi:10.1016/j.ibmb.2015.02.001
110. Tobler A, Nijhout HF (2010) Developmental constraints on the evolution of wing-body allometry in *Manduca sexta*. Evol Dev 12(6):592–600
111. Thaler JS, Contreras H, Davidowitz G (2014) Effects of predation risk and plant resistance on *Manduca sexta* caterpillar feeding behaviour and physiology. Ecol Entomol 39(2):210–216
112. Zhang S, Cao X, He Y, Hartson S, Jiang H (2014) Semi-quantitative analysis of changes in the plasma peptidome of *Manduca sexta* larvae and their correlation with the transcriptome variations upon immune challenge. Insect Biochem Mol Biol 47:46–54
113. Ohshima Y, Suzuki Y (1977) Cloning of the silk fibroin gene and its flanking sequences. Proc Natl Acad Sci U S A 74(12):5363–5367
114. Lecanidou R, Eickbush TH, Rodakis GC, Kafatos FC (1983) Novel B family sequence from an early chorion cDNA library of *Bombyx mori*. Proc Natl Acad Sci U S A 80(7):1955–1959

Chapter 4
Molecular Adaptations of Aphid Biotypes in Overcoming Host-Plant Resistance

Raman Bansal and Andy Michel

Abstract Host-plant resistance (HPR) is a valuable tactic to control pests of agronomic and horticultural crops. Insects are often the most frequent targets of HPR, especially aphids. However, aphids are prone to adapt and overcome this natural pest resistance, which threatens the efficacy, durability, and sustainability of this strategy. In this short review, we focus on recent genetic and molecular biology research that has advanced our mechanistic understanding of aphid biotype evolution with respect to HPR. We highlight studies that have utilized new population genomic, transcriptomic, and metabolomic techniques. We also draw inferences from studies on the evolution of aphid biotype adaptation on different host plants and discuss how these studies can provide a framework to study aphid biotypes. While research shows the existence of multiple, possible routes for overcoming HPR defenses, the exact mechanism(s) remains unclear. An interdisciplinary approach involving multiple fields, including omics research (population and functional genomics, transcriptomics, metabolomics, proteomics, etc.), endosymbiont biology, as well as the ecological interactions between HPR crops and the aphid pests that they target, is needed.

Abbreviations

AFLP	Amplified fragment length polymorphism
Ca^{2+}	Calcium ion
EST	Esterase
GST	Glutathione *S*-transferase
HPR	Host-plant resistance
Hx	Hydroxamic acid
IDE	Inhibitor of digestive enzyme
JA	Jasmonic acid
LRRs	Leucine-rich repeats

R. Bansal • A. Michel (✉)
Department of Entomology, Ohio Agricultural Research and Development Center, The Ohio State University, 1680 Madison Ave, Wooster, OH 44691, USA
e-mail: bansal.67@osu.edu; michel.70@osu.edu

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*, Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_4

MAPK	Mitogen-associated protein kinase
NBS	Nucleotide-binding site
P450	Cytochrome P450 monooxygenase
PSM	Plant secondary metabolite
QTL	Quantitative trait locus
RAPD	Random amplified polymorphic DNA
RFLP	Restriction fragment length polymorphism
RWA	Russian wheat aphid
SA	Salicylic acid
SNP	Single-nucleotide polymorphism

4.1 Host-Plant Resistance and Biotype Evolution

Host-plant resistance (HPR) is a pest management technique that exploits naturally evolved plant defenses for improved and sustainable crop production [1]. Development of crop varieties with resistance to insect and arthropod pests has a long history, starting as early as crop domestication [2–4]. Not only has HPR been implemented as a management tactic for many crops (see [4] for a recent list of 22 crops), it serves as a model to expand our understanding of insect-plant interactions.

Hemipterans are a frequent target for HPR, and perhaps no group is more targeted than aphids—at least 16 crops are bred for aphid resistance [5, 6]. Most of these HPR crops are highly successful, for example, maize and wheat lines developed by traditional plant breeding techniques can limit aphid damage [6–8]. Alfalfa and sorghum with aphid and leafhopper resistance have an annual economic value of over $400 million [6]. In addition to economic savings, HPR promotes the ecological service of biological control [7] and can lead to a decrease in potentially hazardous chemical applications. In some cases, however, the use of HPR in pest management has been challenging and limited due to many factors including efficacy, economic viability, and lack of availability of resistant varieties [8].

Perhaps the most serious challenge to full implementation of HPR crops is their durability in the face of insect biotype evolution [3, 5, 9]. There are several definitions of the term "biotype," but, in the context of agricultural insect pests and HPR interactions, biotypes are specifically defined by their differential survival or fitness on, or adaptation to, host-plant defenses [3, 5]. Insect populations capable of overcoming resistance are considered *virulent* to the HPR plant, whereas those unable to survive and reproduce are referred to as *avirulent* [3]. Furthermore, *compatible interactions* occur when insects can feed and colonize on a plant, as opposed to *incompatible interactions* which result in insect deterrence and/or death.

Insect pests across a variety of taxa have developed biotypes in response to HPR, such as the Hessian fly (*Mayetiola destructor* [10]), the black currant leaf midge (*Dasineura tetensi* [11]), rice brown plant hopper (*Nilaparvata lugens* [12]), and black pine-leaf scale (*Nuculaspis californica* [13]). However, a large portion of documented biotypes are clustered in the order Hemiptera, specifically within the

family Aphididae [2, 3, 5, 14]. Smith and Chuang [6] listed 17 aphid species that have adapted to HPR, and all of these have multiple biotypes, i.e., differential survival with different HPR genes. In some cases it can take years for virulence to evolve—aphid-resistant strawberries were effective for *c.a.* 50 years [6]. Alternatively, biotypes can occur even before the large-scale deployment of HPR, as was the case with the soybean aphid, *Aphis glycines* [9, 15, 16]. The evolution of biotypes in the soybean aphid was a particularly notable example of rapid biotype evolution because, despite the genetic bottleneck during its North American invasion, virulence was observed within 5 years after invasion.

HPR can be a valuable strategy for insect management, but its use as an alternative to insecticides is limited by the evolution of virulent biotypes. Moreover, very few mechanisms of virulent biotype evolution have been described. Understanding the genetic and molecular factors that explain virulence and biotype adaptation is important to develop strategies that limit increases in its frequency, to extend HPR crops' durability and to improve the sustainability of this management tactic. This short review will highlight important advances in our understanding of how virulent biotype adaptation occurs and also show where additional studies and important tools are needed in order to fully use HPR to its potential.

4.2 Population Genomics in Characterizing Biotype Differentiation

Diehl and Bush [17] developed an evolutionarily based framework for the characterization of insect biotypes, largely based on how genetic variation was partitioned among populations. They hypothesized that if biotypes were truly distinct and evolutionarily defined, then greater genetic similarity should exist among individuals of the same biotype rather than between biotypes, i.e., genetic variation would be better explained by biotype designation and not other factors such as geography. However, this framework had not been fully tested until the wide-scale use and practicality of molecular markers enabled such comparisons. Still, most of these early studies focused on population genetics, migration and structure, or comparisons of intraspecific aphid populations on different species of host plants, and, in many of these cases, genetic differentiation supporting host-associated populations were found [18–21]. Yet, there are only a few studies that used these tools to directly compare aphid biotypes on susceptible and resistant cultivars of the same species.

4.2.1 Biotypes of Raspberry Aphids

Raspberry aphids (*Amphorophora idaei* and *A. agathonica*) have several biotypes that are virulent to aphid-resistance genes in raspberry [22–24]. An earlier study using a restriction fragment length polymorphism (RFLP)-based approach to

analyze ribosomal spacer length variability showed discrete patterns among biotype clones. However, when comparing field populations, a greater extent of variability was observed, complicating attempts to associate genetics with specific biotypes [24].

4.2.2 Greenbug Biotypes

The greenbug, *Schizaphis graminum*, is a commonly found aphid of wheat and sorghum in North America which has 8–13 known biotypes [25, 26]. The genetics of greenbug biotypes has been compared using several marker types, including mtDNA sequencing, RFLPs, random amplified polymorphic DNAs (RAPDs), amplified fragment length polymorphisms (AFLPs), and microsatellites [26–29]. All markers largely suggested the presence of three clades, although divergence was relatively recent (0.3–0.6 Ma), and also included polyphyletic assemblages of biotypes; indeed, it appeared that clade representation was better explained by different host-plant species (which included weedy hosts near crop fields) and not by different resistant crop cultivars [28]. The use of 31 microsatellite markers appeared to increase the resolution and ability in defining biotypes, but a population-wide perspective was difficult to interpret as within-biotype variation was not included [26]. Nonetheless, these studies did find substantial genetic variation within aphids from a multitude of wild and cultivated hosts that likely predated the development of resistant cultivars and served as a possible genetic reservoir for adaptation to resistant cultivars [26, 28, 30].

4.2.3 Russian Wheat Aphid Biotypes

The Russian wheat aphid (RWA) (*Diuraphis noxia*) is another significant and worldwide pest of wheat with eight known virulent biotypes in North America [25, 31, 32]. An AFLP comparison of these eight biotypes with other populations from South America, Europe, Africa, and the Middle East showed that at least two invasions occurred in North America, one from Middle East-Africa, and one from Europe [33]. Therefore, North American biotypes did not share recent common ancestors and instead likely emerged after these separate introductions. Two biotypes (RWA1 and 2) were placed in the Middle Eastern-African clade, and biotypes RWA3, 4, and 5 were European in origin. The independent evolution of these biotypes, combined with the high-resolution power of AFLPs, allowed for a clear delineation in genetic differences among biotypes in different clades. However, a comparison of within-biotype variation was not included, as AFLP profiles resulted from pools of 20 individuals. Cui et al. [34] compared sequence polymorphisms of 17 putative salivary transcripts of RWA. A total of 13 of these transcripts contained variation, and some single-nucleotide polymorphisms (SNPs) and indels were

specific to one biotype, albeit at low frequency. Much of the variation was shared among biotypes, and because laboratory strains were used, it is unclear if frequencies in natural populations would have been similar. Interestingly, these authors did find evidence of positive selection and rapid adaptation among these transcripts suggesting that salivary transcripts may play an important role in HPR interactions.

4.2.4 Soybean Aphid Biotypes

A. glycines is a significant agricultural pest of soybean in Asia and is invasive in North America. There are four biotypes, three of which are virulent to various *Rag* genes (**R**esistance to ***A**phis* ***g**lycines*) [15, 16, 35]. Using microsatellites, Michel et al. [36] were able to find diagnostic markers among laboratory colonies of the avirulent biotype 1 and the virulent biotype 2. However, when SNPs were used to compare avirulent (biotype 1) and virulent (biotype 2) *A. glycines* collected from resistant and susceptible plants in the field, no diagnostic markers were found, and genetic differentiation was not apparent among biotypes [37]. These data mirrored the raspberry aphid study [23] in that substantial genetic diversity was found in field populations but did not cluster by biotype. For the soybean aphid, there was a stronger relationship with genetic isolation by geographic distance, aided by large-scale dispersal late in the growing season [38].

While these traditional molecular marker studies have expanded our understanding of biotypic genetic variation, they have not been able to identify aphid genes that may be under natural selection for virulence nor develop reliable diagnostic markers among biotypes. The reasons for this challenge are varied and complicated. From a population-genetic perspective, a molecular marker-based approach for identifying virulence adaptation will be feasible if selection is stronger than gene flow in aphids undergoing full or partial sexual reproduction. In many aphids, this may not be the case. Selection placed on the aphid population by resistant cultivars may lead to the evolution of virulence, but the use of HPR often occurs in a patchwork mosaic, i.e., different cultivars in different areas or in limited quantities (e.g., ~40 % for *Aphis gossypii*-resistant melon, ~50 % for US sorghum, and limited acreage in soybean; see recent reviews [6, 8]). In some cases, like the greenbug, movement may also be to and from wild grasses and other plants [28, 30]. This heterogeneity would result in more balanced polymorphisms instead of the fixed (or nearly fixed) differences needed for frequency-based molecular marker analyses. Sexual reproduction in many of these aphid species allows for recombination, potentially removing any linkage between molecular markers and the virulence gene(s). An additional complication is the often contentious phenotypic definition of biotype [17, 37, 39], which is based on an insect's response to a resistant plant. Insects showing very similar responses can result from very different genetic backgrounds, as was seen with RWA2 and RWA4 [33]. Furthermore, individuals within a biotype designation may not share the same mechanism of virulence and could instead result from

convergent evolution [33] or coadapted gene complexes that provide a more qualitative aspect of virulence [37].

Newer and high-throughput sequencing technologies utilizing whole-genomic approaches have been widely used for characterizing aphid biotypes on different host-plant associations. For example, host-race evolution in the pea aphid, *Acyrthosiphon pisum*, has long been a research focus for understanding insect adaptation and speciation [21, 40, 41]. A recent combination of quantitative trait loci (QTLs) and a genome scan with AFLPs and 137 microsatellites revealed correlated genomic areas under divergent selection among populations on red clover and alfalfa [42]. A different study using 390 microsatellites also found markers under selection among pea aphids from alfalfa, clover, pea [43], and other host plants [44]. When compared to the pea aphid genome [45], 5 of the 11 outlier markers were linked to important genes (two markers related to olfactory and three markers with salivary transcripts). A targeted resequencing array (e.g., exon capture) of known pea aphid genes analyzed sequence variation of 172 loci among pea aphid from alfalfa, clover, and trefoil [46]. Significant genetic differentiation was found among host races, and much of it was focused on receptors for host odor and assessing plant quality (i.e., gustatory). These recent studies in polyphagous host-plant adaptation will certainly serve as foundations for future investigations on virulent biotype evolution to HPR, especially when additional whole-genome and molecular sequence data from multiple aphid species are obtained.

4.3 Molecular Interactions of Aphid Resistance—The Plant Perspective

The molecular interactions between aphids and their host plants have long been a research focus [3, 5, 6, 47]. Despite this research, only two aphid-resistance (R) genes, *Mi-1.2* and *Vat*, have been cloned. *Mi-1.2* in wild tomato confers resistance to the potato aphid, *Macrosiphum euphoribae* [48], and *Vat* in melon provides resistance to *A. gossypii* [49]. Similar to pathogen-resistance genes in plants, both *Mi-1.2* and *Vat* encode proteins containing a nucleotide-binding site (NBS) and leucine-rich repeats (LRRs) [48–52]. *Mi-1.2* guards the aphid-effector target *RME1* in plant cells [53] and activates defense signal transduction as soon as it detects a modification in *RME1*. These plant "R" genes mediate the resistance to aphids through microRNAs [54]. Additional genes in other aphid-resistant plants have been mapped to genic regions known to encode NBS-LRR-like proteins (see [6] for a list of genes).

Molecular studies have shown that the defenses in resistant plants including those possessing NBS-LRR "R" genes are induced after attack by aphids rather than being constitutive [4, 6, 55–59]. The induced defense is advantageous to host plants as it incurs less metabolic cost and is more pest specific [60, 61]. Upon induction, the defense signal is transduced through downstream cascades involving phytohor-

mone pathways which ultimately leads to the synthesis of a variety of defense chemicals (detailed in sections below) [4, 61]. Signaling through the jasmonic acid (JA) pathway seems to be vital for resistance to aphids, although the salicylic acid (SA) pathway can play a role [62]. Interestingly, SA induction by some aphid species makes the plant susceptible, which is seen as a ploy (sometimes called a decoy response) to suppress the more effective JA signaling (detailed below).

4.4 Molecular Interactions of Aphid Resistance—The Aphid Perspective and Virulence Evolution

Upon attack by an avirulent biotype, resistant plants induce defenses, which generally occur in three steps (Fig. 4.1). (1) *Recognition of pest attack*: A plant's surveillance system detects the attack through recognition of the pest's specific signals including molecular patterns and effectors. (2) *Signal transduction*: Detected signals are then carried through a network of signal transduction pathways like mitogen-associated protein kinases (MAPKs) and phytohormones (JA, SA, etc.). (3) *Defensive chemical production*: Signaling pathways eventually lead to the production of plant defense chemicals such as plant secondary metabolites (PSMs), proteases, protease inhibitors, and lectins so as to deter or kill the pest. Virulent biotype adaptation can occur by impeding any of these three steps, such as (1) evading recognition by the plant's surveillance system and preventing defense induction, (2) distorting or manipulating the signal transduction to their own advantage, and (3) developing resistance to plant defense chemicals (Fig. 4.1). Based on the body of knowledge available on aphid biology and aphid-plant interactions, all these scenarios for the defeat of plant resistance by aphid biotypes seem plausible.

4.4.1 Avoidance of Recognition by Plant Surveillance

There is growing evidence which suggests that plants recognize aphid attack through the latter's effector molecules injected into host cells using needle-like stylets [63–66]. Effectors are proteins or other small molecules present in aphid salivary glands which can modify the structure and function of a plant cell [63]. Upon recognition by "R"-gene-mediated surveillance, an aphid effector can trigger the plant defense response. For example, Mp10, an effector from green peach aphid, *Myzus persicae*, induces plant defenses as revealed through its *in planta* transient overexpression in *Nicotiana benthamiana* and activation of both JA and SA signaling pathways [67], which ultimately resulted in reduced aphid fecundity [68]. Similarly, Mp42, Mp56, Mp57, Mp58, and other effectors from the green peach aphid are thought to induce plant defenses, as their transient overexpression in *N. tabacum* and *Arabidopsis thaliana* caused a reduction in aphid fecundity [68, 69].

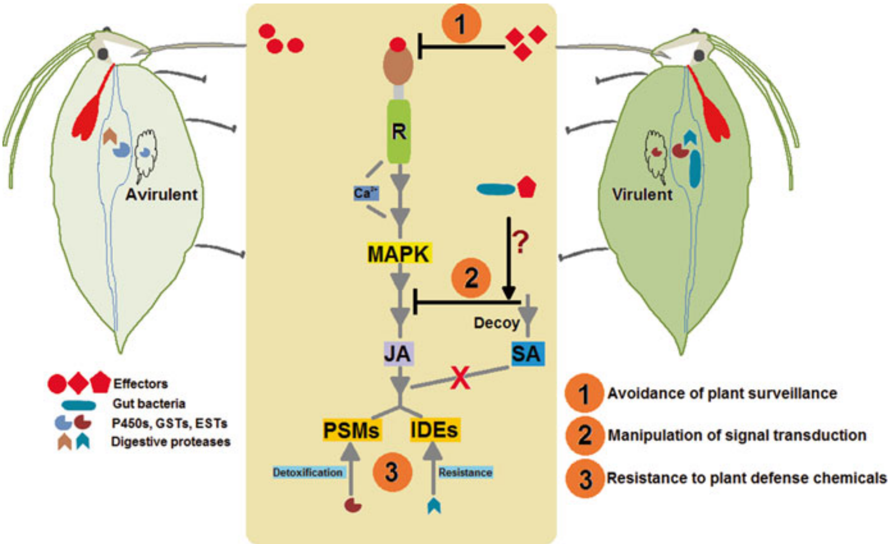


Fig. 4.1 A model summarizing putative strategies adopted by virulent aphid biotypes to overcome HPR. *1. Avoidance of plant surveillance*: Virulent biotypes can avoid recognition by a plant R (resistance)-gene-mediated guard system through diversified/novel salivary effectors. *2. Manipulation of signal transduction*: Virulent biotypes can distort and/or hinder signal transduction pathways to either block the signal transduction or manipulate and divert the transduction in such a way so that the signal is not transduced properly. The latter scenario can lead to an ineffective SA pathway in place of a biologically potent JA pathway, referred to as a decoy response. Aphid gut bacteria or salivary effectors could be involved in inducing the decoy response. *3. Resistance to plant defense chemicals*: Virulent biotypes may evolve resistance to plant secondary metabolites (PSMs) through a detoxification system comprised of cytochrome P450s (P450s), glutathione-s-transferases (GSTs), and esterases (ESTs). Usually, gut and fat body are sites for the occurrence of detoxification events within insects. Virulent biotypes may evolve resistance to inhibitors of digestive enzymes (IDEs) produced by plants through various strategies as described in the text. Ca^{2+} are also involved in other cellular activities during stress such as production of reactive oxygen species (not indicated here). Plant signaling events shown here are based on those described in Wu and Baldwin [61]. *MAPK* mitogen-associated protein kinases, *JA* jasmonic acid, *SA* salicylic acid, Ca^{2+} calcium ions

However, from the aphid's perspective, effectors are produced and secreted into plant cells not to induce plant defenses, but to promote their own virulence and colonization [63]. Indeed, effectors like C002 (from the pea aphid and the green peach aphid) [70, 71], Mp1, Mp2, Mp55 (from *M. persicae*) [69, 71], and Me10 and Me23 (from the potato aphid) [72] increase aphid fecundity and virulence on their respective host plants. Thus, to successfully colonize and adapt on resistant plants, virulent aphids may keep the plant defenses in an un-induced state by evading the plant's surveillance through employing a diversified effector or altogether abandoning a particular effector [63] (Fig. 4.1). Population genomics and proteomics research suggests the adoption of diversified effectors as a possible strategy as evidenced by a strong positive selection in many effector genes [34, 43, 71, 73]. For example,

higher non-synonymous variations exist among salivary effector transcripts of biotypes in the RWA and the pea aphid [34, 43]. Future research on the comparative functional analysis of diversified effectors from disparate biotypes within and among aphid species would improve our ability to understand the role of effectors in virulent biotype adaptation.

4.4.2 Manipulation of Signal Transduction Pathways

Once an aphid attack is recognized, the signal is transduced through various cellular compartments to meet the end goal of producing defense chemicals. Signal transduction occurs through multiple layers of intracellular transduction pathways involving Ca^{2+}, reactive oxygen species, MAPKs, phytohormones, and transcription factors [61]. Theoretically, virulent biotypes may either distort and/or hinder any of these pathways to either block the signal transduction or manipulate and divert the transduction in such a way so that the signal is not transduced properly (a "decoy" response; see below) (Fig. 4.1). Both these scenarios will prevent the synthesis of desired defense toxins.

In most cases, JA signaling is the key phytohormone pathway that suppresses aphid colonization [74]. This was made evident through research on the model plant *Arabidopsis* when mutants deficient in JA signaling lost resistance to aphids [75, 76], whereas mutants that were compromised for SA signaling retained resistance to aphids [55, 77]. Furthermore, in compatible interactions where aphids could successfully colonize the plants, higher SA-pathway transcripts were observed; on the other hand, reduced or slightly increased JA-pathway transcripts were detected in susceptible plants [55, 75, 78–80]. These observations have led to the "decoy" hypothesis where aphids are believed to hijack plant defense signaling by manipulating the signal transduction away from the biologically potent JA pathway and toward the ineffective SA pathway. The often negative cross talk of SA signaling with JA signaling can also hinder the effective deployment of plant resistance to aphids. Though the mechanism of manipulation of phytohormone signaling by aphids is not well understood, the infection-promoting effectors or insect gut bacteria could be involved [81] (discussed below). Future studies on the comparative transcriptomic and biochemical analysis of phytohormone and other signaling constituents in cultivars infested with virulent and avirulent biotypes will help to shed light on aphid biotype evolution through manipulation of plant defense signaling.

4.4.3 Development of Resistance to Plant Defense Chemicals

In plants, the successful transduction of induced signal leads to production of a multifaceted defense that can be broadly categorized as toxic and anti-nutritious [60]. Both toxicity and anti-nutrition are manifested through a variety of plant defense chemicals such as PSMs, proteases, and protease inhibitors.

4.4.3.1 Resistance to PSMs

PSMs are metabolic by-products which are not required for normal plant growth and development but possess direct toxicity to pests including aphids [82]. To counter PSMs, insects have evolved their own defense strategies, typically involving detoxification enzymes such as cytochrome P450 monooxygenases (P450s), glutathione *S*-transferases (GSTs), and esterases (ESTs). These detoxification enzymes can readily metabolize PSMs to limit their effectiveness (Fig. 4.1) and lead to virulent biotype adaptation. PSM resistance has been reported in numerous insects [83–86]. Employing detoxification enzymes incurs a low energy and fitness costs to aphids, thus makes it a favorable strategy to overcome HPR [87]. The role of P450s in mediating biotypic host-plant adaptation is supported by the fact that generalist aphids carry a significantly larger repertoire of P450 enzymes than specialists. For example, the generalist green peach aphid, which feeds on more than 100 species in 40 different plant families, has at least 40 % more P450 genes compared to the pea aphid, a specialist which feeds only on a few species within a single plant family (Fabaceae) [88].

There are at least two possible ways by which detoxification can lead to PSM resistance and biotype adaptation: (1) The detoxification enzyme can be produced in higher amounts, most likely through overexpression. For example, the green peach aphid adapts to glucosinolates (a family of PSMs) in *Sinapis alba* by producing more GSTs compared to when feeding on glucosinolate-free *Vicia faba* [83]. (2) Mutation(s) can occur in the catalytic site of detoxification enzymes enabling a much more efficient and effective neutralization of the PSM. In fact, detoxification genes are induced in avirulent aphid biotypes when fed with resistant plants or are exposed to PSMs. For example, GSTs have higher expression in the avirulent biotype 1 of RWA fed with wheat plants containing the *Dn4* resistance gene [89]. Higher activity of GSTs and ESTs has been found in the cereal aphid, *Sitobion avenae*, after feeding on resistant wheat with high concentrations of phenolics (PSMs) or when exposed to gramine (an alkaloid PSM) [90, 91]. Similarly, higher enzymatic activities for P450s, GSTs, and ESTs were found in cereal aphid fed with hydroxamic acid (Hx)-containing wheat compared to those fed with Hx-free oats [92]. Higher EST activity occurred in the corn aphid, *Rhopalosiphum padi*, when feeding on resistant wheat compared to those fed with susceptible plants [93, 94]. Similarly, certain P450s, GSTs, and ESTs are induced when the avirulent biotype 1 of the soybean aphid feeds on a soybean plant possessing the *Rag1* resistance gene [95].

4.4.3.2 Resistance to Inhibitors of Digestive Enzymes

As a part of their defense, plants induce the production of inhibitors that target insect digestive enzymes, the majority of which are proteases and amylases [96]. Protease inhibitors targeting various aphid species have been characterized in different plants [97–99]. However, insects are known to adapt to plant protease

inhibitors in a number of ways (Fig. 4.1): (1) inactivation of protease inhibitors by direct proteolysis by insect gut proteinases [100, 101], (2) overproduction of existing digestive proteases [102], (3) expression of inhibitor-insensitive proteases [101, 103, 104], and (4) inducing isoforms of inhibitor-sensitive proteases [105]. The latter three strategies essentially result in redeployment of the insect digestive arsenal which is regulated by the alteration in gene expression of different digestive enzymes. Indeed, like in other insects, aphids show differential expression of gut digestive enzymes when feeding on HPR crops. For example, there is significant differential regulation of gut proteases among the avirulent biotype 1 and the virulent biotype 2 of RWA fed with *Dn4* wheat [89]. Similarly, there is significantly differential regulation of protease and protease inhibitors in the virulent biotype 3 and avirulent biotype 1 of the soybean aphid feeding on resistant (*Rag1*) soybean [95]. Moreover, aphids exhibit a massive expansion in their repertoire of cathepsin B genes, the major digestive proteinases of hemipterans which can overcome plant protease inhibitors [106].

4.4.4 Role of Bacterial Symbionts in Aphid Biotype Evolution

Aphids are well known for their symbiotic relationships with bacteria. The pea aphid is known to harbor three kinds of bacteria: (1) the obligate endosymbiont, *Buchnera*; (2) several facultative endosymbionts (*Hamiltonella*, *Regiella*, *Serratia*, *Rickettsia*, and *Spiroplasma*); and (3) extracellular gut microbiota which reside in the lumen of the digestive tract (e.g., *Pantoea*, *Bacillus*). However, the contribution of endosymbionts for virulent biotype adaptation may be limited because their intracellular lifestyle hampers the release of factors or gene products directly into the salivary secretions or the gut lumen, where they might assist in overcoming host-plant recognition/defenses, detoxifying plant defense chemicals, or improving digestion [107].

The phenomenon of biotype evolution is characterized by a perpetual arms race requiring defense against novel challenges posed by host plants. Therefore, *Buchnera* is highly unlikely to be involved in aphid biotype evolution as it possesses a dramatically reduced genome and does not acquire novel genes in its symbiotic relationship with aphids [107, 108]. In fact, recent studies suggest *Buchnera* might actually be an antagonist, though inadvertently, to its host aphid. *Buchnera*'s chaperonin, GroEL, can act as a molecular pattern to trigger a plant's defense response, which can negatively affect aphid growth and fecundity [69, 109]. It is speculated that aphids have evolved effectors to suppress *Buchnera* GroEL-triggered immunity in host plants [109]. Some caution is required in this interpretation, however, because these results are based on *in planta* overexpression or exogenous application of GroEL. Alternatively, *Buchnera* may be involved in greenbug virulence [110]. Proteomic variation linked with unique sequence polymorphisms in the EF-Tu protein from *Buchnera* was found within the highly virulent biotype H when compared to avirulent biotypes, although the exact mechanism or role of this

protein is unclear [110]. Future improvements on in vivo studies involving *Buchnera* and research on the localization of *Buchnera* proteins in cells of aphid-infested plants will better discern *Buchnera*'s direct role in virulent biotype evolution.

There has been some evidence to suggest that facultative endosymbionts drive aphid biotype specialization. For example, *Regiella insecticola* improved the fitness of the pea aphid on white clover but not on vetch plants [111]; however, subsequent studies did not support these results [112–114]. In another study, particular facultative endosymbiont species were found to be associated with a particular host-specialized biotype of *A. pisum* [115]. Nonetheless, inferences drawn from such surveys on the role of facultative endosymbionts in governing aphid biotype evolution could be misleading, and, to date, there are no studies to suggest that these bacteria play a role for virulent biotype evolution in relation to HPR. The association of a facultative endosymbiont with a particular aphid biotype may occur due to many other factors, as discussed in [107].

Due to their location, gut bacteria are perhaps best situated to play a direct role in aphid biotype evolution. Bacteria such as *Staphylococcus*, *Pseudomonas*, *Acinetobacter*, *Pantoea*, *Bacillus*, and *Brevundimonas* have been detected from aphid gut; however, functions for most of these are not well known [116–119]. In general, insect gut bacteria are known to perform three major activities which may be significant in virulent biotype adaptation: (1) They aid in digestion by producing inhibitor-resistant proteases [120]; (2) insect gut bacteria can detoxify PSMs, the major plant defense chemicals inside insect gut [121]; and (3) insect gut bacteria can induce decoy responses to suppress effective host-plant defense [81]. Conversely, there is a lack of concrete evidence on any role of gut bacteria in virulent biotype adaptation in aphids, with a few studies reporting only the diversity and abundance of an insect's gut and their transient presence occurring due to inconsistent infections [119]. Nonetheless, there is data to support a role for symbionts in virulent biotype adaptation to HPR in plant hoppers and leafhoppers (suborder Auchenorrhyncha), which are close relatives of the Aphididae (suborder Sternorrhyncha) [122]. More research is certainly necessary regarding the possibility of endosymbionts contributing to virulent biotype adaptation in aphids.

4.5 Conclusion

As much investment is made toward the development of HPR crops, it is imperative to understand virulent biotype adaptation and improve management strategies that extend their durability. There has been an increase in population genomics and transcriptomics research on biotype adaptation to diverse host plants which can offer clues for virulent biotype evolution. Although similar, these comparisons may also be functionally different, as the mechanism(s) of virulent biotype adaptation to HPR (sources within a single plant species) may involve greater specificity than adapting to different plant species. Transitioning among host plants in different genera may involve more complicated and diverse adaptations, stronger selection pressures, and

reproductive isolation. These factors may not all exist to the same degree during adaptation to HPR plants or varieties which may differ by one or a few genes. Since the success of large-scale population genomics and transcriptomics depends on the strength of the selective footprint [123–125], these approaches alone may not be enough to reveal mechanisms of virulent biotype adaptation, i.e., the genomic islands of divergence are too few and too small to be detected from the sea of neutral variation [126, 127]. A combined approach will be necessary which not only includes proteomics and metabolomics but also the interactions with the incredible microbial diversity that aphids house as well. Research using natural populations is also needed to capture the extent of genetic variation and diversity in both aphids and bacteria, as studies have shown dramatic differences in laboratory and natural populations in both taxa [36, 119]. Virulent biotype adaptation to HPR will also need to be investigated in the context of complex agroecosystems that include natural enemies, insecticides, and a patchwork mosaic of crop varieties that differ in maturity, nutrition (e.g., oil, protein, sugars), and many other phenotypic traits. Expanding our understanding of virulent biotype adaptation in aphids will help maintain the efficacy of current HPR crops and will also provide a solid foundation to study virulence or resistance development when the next generation of RNA interference-based insect control is implemented [128].

Acknowledgments We would like to thank members of the Michel Laboratory including L. Wallace and J. Wenger, the M.A.R Mian laboratory at USDA-ARS, and the SoyRes Team from the Center of Applied Plant Sciences at The Ohio State University. Support for this work was provided by the Ohio Agricultural Research and Development Center, The Ohio State University, as well as various soybean checkoff organizations, including the Ohio Soybean Council, North Central Soybean Research Program, and the United Soybean Board.

References

1. Painter RH (1951) Insect resistance in crop plants. Macmillan, New York
2. Panda N, Khush GS (1995) Host plant resistance to insects. CAB International, Wallingford
3. Smith CM (2005) Plant resistance to arthropods: molecular and conventional approaches. Springer, Dordrecht
4. Smith CM, Clement SL (2012) Molecular bases of plant resistance to arthropods. Annu Rev Entomol 57(1):309–328
5. Van Emden HF, Harrington R (2007) Aphids as crop pests. In: van Emden HF, Richard Harrington R (eds). CABI Publishing
6. Smith CM, Chuang W-P (2014) Plant resistance to aphid feeding: behavioral, physiological, genetic and molecular cues regulate aphid host selection and feeding. Pest Manag Sci 70(4):528–540
7. McCarville MT, O'Neal ME (2012) Measuring the benefit of biological control for single gene and pyramided host plant resistance for *Aphis glycines* (Hemiptera: Aphididae) Management. J Econ Entomol 105(5):1835–1843
8. Hesler LS, Chiozza MV, O'Neal ME, MacIntosh GC, Tilmon KJ, Chandrasena DI, Tinsley NA, Cianzio SR, Costamagna AC, Cullen EM, DiFonzo CD, Potter BD, Ragsdale DW,

Steffey K, Koehler KJ (2013) Performance and prospects of *Rag* genes for management of soybean aphid. Entomol Exp Appl 147(3):201–216

9. Michel A, Omprakah M, Mian R (2011) Evolution of soybean aphid biotypes: understanding and managing virulence to host-plant resistance. In: Sudaric A (ed) Soybean – molecular aspects of breeding. InTech, pp 355–372. doi:10.5772/14407
10. Ratcliffe RH, Cambron SE, Flanders KL, Bosque-Perez NA, Clement SL, Ohm HW (2000) Biotype composition of Hessian fly (Diptera: Cecidomyiidae) populations from the southeastern, midwestern, and northwestern United States and virulence to resistance genes in wheat. J Econ Entomol 93(4):1319–1328
11. Hellqvist S (2001) Biotypes of *Dasineura tetensi*, differing in ability to gall and develop on black currant genotypes. Entomol Exp Appl 98(1):85–94
12. Sōgawa K (1982) The rice brown planthopper: feeding physiology and host plant interactions. Annu Rev Entomol 27(1):49–73
13. Edmunds GF, Alstad DN (1978) Coevolution in insect herbivores and conifers. Science 199(4332):941–945
14. Blackman RL, Eastop VF (1984) Aphids on the world's crops. An identification and information guide. Wiley, Chichester
15. Kim KS, Hill CB, Hartman GL, Mian MAR, Diers BW (2008) Discovery of soybean aphid biotypes. Crop Sci 48(3):923–928
16. Hill CB, Crull L, Herman TK, Voegtlin DJ, Hartman GL (2010) A new soybean aphid (Hemiptera: Aphididae) biotype identified. J Econ Entomol 103(2):509–515
17. Diehl SR, Bush GL (1984) An evolutionary and applied perspective of insect biotypes. Annu Rev Entomol 29(1):471–504
18. Sunnucks P, De Barro PJ, Lushai G, Maclean N, Hales D (1997) Genetic structure of an aphid studied using microsatellites: cyclic parthenogenesis, differentiated lineages and host specialization. Mol Ecol 6(11):1059–1073
19. Lushai G, Markovitch O, Loxdale HD (2002) Host-based genotype variation in insects revisited. Bull Entomol Res 92(2):159–164
20. Lozier JD, Roderick GK, Mills NJ (2009) Tracing the invasion history of mealy plum aphid, *Hyalopterus pruni* (Hemiptera: Aphididae), in North America: a population genetics approach. Biol Invasions 11(2):299–314
21. Peccoud J, Ollivier A, Plantegenest M, Simon J-C (2009) A continuum of genetic divergence from sympatric host races to species in the pea aphid complex. Proc Natl Acad Sci 106(18):7495–7500
22. Converse RH, Daubeny HA, Stace-Smith R, Russell LM, Koch EJ, Wiggans SC (1971) Search for biological races in *Amphorophora agathonica* Hottes on red raspberries. Can J Plant Sci 51(2):81–85
23. Birch ANE, Fenton B, Malloch G, Jones AT, Phillips MS, Harrower BE, Woodford JAT, Catley MA (1994) Ribosomal spacer length variability in the large raspberry aphid, *Amphorophora idaei* (Aphidinae: Macrosiphini). Insect Mol Biol 3(4):239–245
24. Dossett M, Kempler C (2012) Biotypic diversity and resistance to the raspberry aphid *Amphorophora agathonica* in Pacific Northwestern North America. J Am Soc Hortic Sci 137(6):445–451
25. Burd JD, Porter DR, Puterka GJ, Haley SD, Peairs FB (2006) Biotypic variation among north American Russian wheat aphid (Homoptera: Aphididae) populations. J Econ Entomol 99(5):1862–1866
26. Weng Y, Perumal A, Burd JD, Rudd JC (2010) Biotypic diversity in Greenbug (Hemiptera: Aphididae): microsatellite-based regional divergence and host-adapted differentiation. J Econ Entomol 103(4):1454–1463
27. Shufran KA, Burd JD, Anstead JA, Lushai G (2000) Mitochondrial DNA sequence divergence among greenbug (Homoptera: Aphididae) biotypes: evidence for host-adapted races. Insect Mol Biol 9(2):179–184

28. Anstead JA, Burd JD, Shufran KA (2002) Mitochondrial DNA sequence divergence among *Schizaphis graminum* (Hemiptera: Aphididae) clones from cultivated and non-cultivated hosts: haplotype and host associations. Bull Entomol Res 92(1):17–24
29. Zhu-Salzman K, Li H, Klein PE, Gorena RL, Salzman RA (2003) Using high-throughput amplified fragment length polymorphism to distinguish sorghum greenbug (Homoptera: Aphididae) biotypes. Agric For Entomol 5(4):311–315
30. Anstead JA, Burd JD, Shufran KA (2003) Over-summering and biotypic diversity of *Schizaphis graminum* (Homoptera: Aphididae) populations on noncultivated grass hosts. Environ Entomol 32(3):662–667
31. Haley SD, Peairs FB, Walker CB, Rudolph JB, Randolph TL (2004) Occurrence of a new Russian wheat aphid biotype in Colorado. Crop Sci 44(5):1589
32. Weiland AA, Peairs FB, Randolph TL, Rudolph JB, Haley SD, Puterka GJ (2008) Biotypic diversity in Colorado Russian wheat aphid (Hemiptera: Aphididae) populations. J Econ Entomol 101(2):569–574
33. Liu X, Marshall JL, Stary P, Edwards O, Puterka G, Dolatti L, El Bouhssini M, Malinga J, Lage J, Smith CM (2010) Global phylogenetics of *Diuraphis noxia* (Hemiptera: Aphididae), an invasive aphid species: evidence for multiple invasions into North America. J Econ Entomol 103(3):958–965
34. Cui F, Michael Smith C, Reese J, Edwards O, Reeck G (2012) Polymorphisms in salivary-gland transcripts of Russian wheat aphid biotypes 1 and 2. Insect Sci 19(4):429–440
35. Alt J, Ryan-Mahmutagic M (2013) Soybean aphid biotype 4 identified. Crop Sci 53(4):1491–1495
36. Michel AP, Zhang W, Mian MAR (2010) Genetic diversity and differentiation among laboratory and field populations of the soybean aphid, *Aphis glycines*. Bull Entomol Res 100(06):727–734
37. Wenger JA, Michel AP (2013) Implementing an evolutionary framework for understanding genetic relationships of phenotypically defined insect biotypes in the invasive soybean aphid (*Aphis glycines*). Evol Appl 6:1041–1053
38. Orantes LC, Zhang W, Mian MAR, Michel AP (2012) Maintaining genetic diversity and population panmixia through dispersal and not gene flow in a holocyclic heteroecious aphid species. Heredity (Edinb) 109:127–134
39. Downie DA (2010) Baubles, bangles, and biotypes: a critical review of the use and abuse of the biotype concept. J Insect Sci 10:1–18. doi:http://dx.doi.org/10.1673/031.010.14136
40. Via S (1999) Reproductive isolation between sympatric races of pea aphids. I. Gene flow restriction and habitat choice. Evolution 53(5):1446–1457
41. Ferrari J, Via S, Godfray HCJ (2008) Population differentiation and genetic variation in performance on eight hosts in the pea aphid complex. Evolution 62(10):2508–2524
42. Via S, Conte G, Mason-Foley C, Mills K (2012) Localizing F(ST) outliers on a QTL map reveals evidence for large genomic regions of reduced gene exchange during speciation-with-gene-flow. Mol Ecol 21(22):5546–5560
43. Jaquiéry J, Stoeckel S, Nouhaud P, Mieuzet L, Mahéo F, Legeai F, Bernard N, Bonvoisin A, Vitalis R, Simon J-C (2012) Genome scans reveal candidate regions involved in the adaptation to host plant in the pea aphid complex. Mol Ecol 21(21):5251–5264
44. Nouhaud P, Peccoud J, Mahéo F, Mieuzet L, Jaquiéry J, Simon J-C (2014) Genomic regions repeatedly involved in divergence among plant-specialized pea aphid biotypes. J Evol Biol 27:2013–2020
45. International Aphid Genomics Consortium (2010) Genome sequence of the pea aphid *Acyrthosiphon pisum*. PLoS Biol 8(2):e1000313
46. Smadja CM, Canbäck B, Vitalis R, Gautier M, Ferrari J, Zhou J-J, Butlin RK (2012) Large-scale candidate gene scan reveals the role of chemoreceptor genes in host plant specialization and speciation in the pea aphid. Evolution 66(9):2723–2738
47. Smith CM, Boyko EV (2007) The molecular bases of plant resistance and defense responses to aphid feeding: current status. Entomol Exp Appl 122(1):1–16

48. Rossi M, Goggin FL, Milligan SB, Kaloshian I, Ullman DE, Williamson VM (1998) The nematode resistance gene *Mi* of tomato confers resistance against the potato aphid. Proc Natl Acad Sci U S A 95(17):9750–9754
49. Brotman Y, Silberstein L, Kovalski I, Perin C, Dogimont C, Pitrat M, Klingler J, Thompson A, Perl-Treves R (2002) Resistance gene homologues in melon are linked to genetic loci conferring disease and pest resistance. Theor Appl Genet 104(6–7):1055–1063
50. Kaloshian I, Kinsey MG, Ullman DE, Williamson VM (1997) The impact of Meu1-mediated resistance in tomato on longevity, fecundity and behavior of the potato aphid, *Macrosiphum euphorbiae*. Entomol Exp Appl 83(2):181–187
51. Milligan SB, Bodeau J, Yaghoobi J, Kaloshian I, Zabel P, Williamson VM (1998) The root knot nematode resistance gene Mi from tomato is a member of the leucine zipper, nucleotide binding, leucine-rich repeat family of plant genes. Plant Cell 10(8):1307–1319
52. Goggin FL, Jia L, Shah G, Hebert S, Williamson VM, Ullman DE (2006) Heterologous expression of the Mi-1.2 gene from tomato confers resistance against nematodes but not aphids in eggplant. Mol Plant Microbe Interact 19(4):383–388
53. Bhattarai KK, Li Q, Liu Y, Dinesh-Kumar SP, Kaloshian I (2007) The MI-1-mediated pest resistance requires Hsp90 and Sgt1. Plant Physiol 144(1):312–323
54. Sattar S, Song Y, Anstead JA, Sunkar R, Thompson GA (2012) Cucumis melo microRNA expression profile during aphid herbivory in a resistant and susceptible interaction. Mol Plant Microbe Interact 25(6):839–848
55. Moran PJ, Thompson GA (2001) Molecular responses to aphid feeding in Arabidopsis in relation to plant defense pathways. Plant Physiol 125(2):1074–1085
56. Coppola V, Coppola M, Rocco M, Digilio MC, D'Ambrosio C, Renzone G, Martinelli R, Scaloni A, Pennacchio F, Rao R, Corrado G (2013) Transcriptomic and proteomic analysis of a compatible tomato-aphid interaction reveals a predominant salicylic acid-dependent plant response. BMC Genomics 14:515
57. Kuśnierczyk A, Tran DHT, Winge P, Jørstad TS, Reese JC, Troczyńska J, Bones AM (2011) Testing the importance of jasmonate signalling in induction of plant defences upon cabbage aphid (*Brevicoryne brassicae*) attack. BMC Genomics 12:423
58. Li Y, Zou J, Li M, Bilgin DD, Vodkin LO, Hartman GL, Clough SJ (2008) Soybean defense responses to the soybean aphid. New Phytol 179(1):185–195
59. McHale L, Tan X, Koehl P, Michelmore RW (2006) Plant NBS-LRR proteins: adaptable guards. Genome Biol 7(4):212
60. Chen M-S (2008) Inducible direct plant defense against insect herbivores: a review. Insect Sci 15(2):101–114
61. Wu J, Baldwin IT (2010) New insights into plant responses to the attack from insect herbivores. Annu Rev Genet 44:1–24
62. Donovan MP, Nabity PD, DeLucia EH (2012) Salicylic acid-mediated reductions in yield in *Nicotiana attenuata* challenged by aphid herbivory. Arthropod Plant Interact 7(1):45–52
63. Hogenhout SA, Bos JIB (2011) Effector proteins that modulate plant–insect interactions. Curr Opin Plant Biol 14(4):422–428
64. Hogenhout SA, Van der Hoorn RAL, Terauchi R, Kamoun S (2009) Emerging concepts in effector biology of plant-associated organisms. Mol Plant Microbe Interact 22(2):115–122
65. Elzinga DA, Jander G (2013) The role of protein effectors in plant-aphid interactions. Curr Opin Plant Biol 16(4):451–456
66. Rodriguez PA, Bos JIB (2013) Toward understanding the role of aphid effectors in plant infestation. Mol Plant Microbe Interact 26(1):25–30
67. Rodriguez PA, Stam R, Warbroek T, Bos JIB (2014) Mp10 and Mp42 from the aphid species *Myzus persicae* trigger plant defenses in *Nicotiana benthamiana* through different activities. Mol Plant Microbe Interact 27(1):30–39
68. Bos JIB, Prince D, Pitino M, Maffei ME, Win J, Hogenhout SA (2010) A functional genomics approach identifies candidate effectors from the aphid species *Myzus persicae* (green peach aphid). PLoS Genet 6(11):e1001216

69. Elzinga DA, de Vos M, Jander G (2014) Suppression of plant defenses by a *Myzus persicae* (green peach aphid) salivary effector protein. Mol Plant Microbe Interact 27(7):747–s756
70. Mutti NS, Louis J, Pappan LK, Pappan K, Begum K, Chen M-S, Park Y, Dittmer N, Marshall J, Reese JC, Reeck GR (2008) A protein from the salivary glands of the pea aphid, *Acyrthosiphon pisum*, is essential in feeding on a host plant. Proc Natl Acad Sci U S A 105(29):9965–9969
71. Pitino M, Hogenhout SA (2013) Aphid protein effectors promote aphid colonization in a plant species-specific manner. Mol Plant Microbe Interact 26(1):130–139
72. Atamian HS, Chaudhary R, Cin VD, Bao E, Girke T, Kaloshian I (2013) In planta expression or delivery of potato aphid *Macrosiphum euphorbiae* effectors Me10 and Me23 enhances aphid fecundity. Mol Plant Microbe Interact 26(1):67–74
73. Carolan JC, Caragea D, Reardon KT, Mutti NS, Dittmer N, Pappan K, Cui F, Castaneto M, Poulain J, Dossat C, Tagu D, Reese JC, Reeck GR, Wilkinson TL, Edwards OR (2011) Predicted effector molecules in the salivary secretome of the pea aphid (*Acyrthosiphon pisum*): a dual transcriptomic/proteomic approach. J Proteome Res 10(4):1505–1518
74. Walling LL (2008) Avoiding effective defenses: strategies employed by phloem-feeding insects. Plant Physiol 146(3):859–866
75. Ellis C, Karafyllidis I, Turner JG (2002) Constitutive activation of jasmonate signaling in an Arabidopsis mutant correlates with enhanced resistance to *Erysiphe cichoracearum*, *Pseudomonas syringae*, and *Myzus persicae*. Mol Plant Microbe Interact 15(10):1025–1030
76. Mewis I, Tokuhisa JG, Schultz JC, Appel HM, Ulrichs C, Gershenzon J (2006) Gene expression and glucosinolate accumulation in *Arabidopsis thaliana* in response to generalist and specialist herbivores of different feeding guilds and the role of defense signaling pathways. Phytochemistry 67(22):2450–2462
77. Pegadaraju V, Knepper C, Reese J, Shah J (2005) Premature leaf senescence modulated by the Arabidopsis PHYTOALEXIN DEFICIENT4 gene is associated with defense against the phloem-feeding green peach aphid. Plant Physiol 139(4):1927–1934
78. Zhu-Salzman K, Salzman RA, Ahn J-E, Koiwa H (2004) Transcriptional regulation of sorghum defense determinants against a phloem-feeding aphid. Plant Physiol 134(1):420–431
79. De Vos M, Van Oosten VR, Van Poecke RMP, Van Pelt JA, Pozo MJ, Mueller MJ, Buchala AJ, Métraux J-P, Van Loon LC, Dicke M, Pieterse CMJ (2005) Signal signature and transcriptome changes of Arabidopsis during pathogen and insect attack. Mol Plant Microbe Interact 18(9):923–937
80. Gao L-L, Anderson JP, Klingler JP, Nair RM, Edwards OR, Singh KB (2007) Involvement of the octadecanoid pathway in bluegreen aphid resistance in *Medicago truncatula*. Mol Plant Microbe Interact 20(1):82–93
81. Chung SH, Rosa C, Scully ED, Peiffer M, Tooker JF, Hoover K, Luthe DS, Felton GW (2013) Herbivore exploits orally secreted bacteria to suppress plant defenses. Proc Natl Acad Sci U S A 110(39):15728–15733
82. Kessler A, Baldwin IT (2002) Plant responses to insect herbivory: the emerging molecular analysis. *Annu. Rev*. Plant Biol 53:299–328
83. Francis F, Vanhaelen N, Haubruge E (2005) Glutathione S-transferases in the adaptation to plant secondary metabolites in the *Myzus persicae* aphid. Arch Insect Biochem Physiol 58(3):166–174
84. Zhang M, Fang T, Pu G, Sun X, Zhou X, Cai Q (2013) Xenobiotic metabolism of plant secondary compounds in the English grain aphid, *Sitobion avenae* (F.) (Hemiptera: Aphididae). Pestic Biochem Physiol 107(1):44–49
85. Després L, David J-P, Gallet C (2007) The evolutionary ecology of insect resistance to plant chemicals. Trends Ecol Evol 22(6):298–307
86. Li X, Schuler MA, Berenbaum MR (2007) Molecular mechanisms of metabolic resistance to synthetic and natural xenobiotics. Annu Rev Entomol 52:231–253
87. Castañeda LE, Figueroa CC, Fuentes-Contreras E, Niemeyer HM, Nespolo RF (2009) Energetic costs of detoxification systems in herbivores feeding on chemically defended host

plants: a correlational study in the grain aphid, *Sitobion avenae*. J Exp Biol 212(Pt 8):1185–1190
88. Ramsey JS, Rider DS, Walsh TK, De Vos M, Gordon KHJ, Ponnala L, Macmil SL, Roe BA, Jander G (2010) Comparative analysis of detoxification enzymes in *Acyrthosiphon pisum* and *Myzus persicae*. Insect Mol Biol 19(Suppl 2):155–164
89. Anathakrishnan R, Sinha DK, Murugan M, Zhu KY, Chen M-S, Zhu YC, Smith CM (2014) Comparative gut transcriptome analysis reveals differences between virulent and avirulent Russian wheat aphids, *Diuraphis noxia*. Arthropod Plant Interact 8(2):79–88
90. Leszczynski B, Urbanska A, Matok H, Dixon AFG (1993) Detoxifying enzymes of the grain aphide. Bull OILB SROP 16:165–172
91. Cai Q-N, Han Y, Cao Y-Z, Hu Y, Zhao X, Bi J-L (2009) Detoxification of gramine by the cereal aphid *Sitobion avenae*. J Chem Ecol 35(3):320–325
92. Loayza-Muro R, Figueroa CC, Niemeyer HM (2000) Effect of two wheat cultivars differing in hydroxamic acid concentration on detoxification metabolism in the aphid *Sitobion avenae*. J Chem Ecol 26(12):2725–2736
93. Chen J, Song D, Cai C, Cheng D, Tian Z (1997) Biochemical studies on wheat resistance to the grain aphid, *Rhopalosiphum padi* (L.). Acta Entomol Sin 40:186–189
94. Cai QN, Zhang QW, Cheo M (2004) Contribution of indole alkaloids to *Sitobion avenae* (F.) resistance in wheat. J Appl Entomol 128(8):517–521
95. Bansal R, Mian M, Mittapalli O, Michel AP (2014) RNA-Seq reveals a xenobiotic stress response in the soybean aphid, *Aphis glycines*, when fed aphid-resistant soybean. BMC Genomics 15(1):972
96. Habib H, Fazili KM (2007) Plant protease inhibitors: a defense strategy in plants. Biotechnol Mol Biol Rev 2(3):68–85
97. Ceci LR, Volpicella M, Rahbé Y, Gallerani R, Beekwilder J, Jongsma MA (2003) Selection by phage display of a variant mustard trypsin inhibitor toxic against aphids. Plant J 33(3):557–566
98. Azzouz H, Cherqui A, Campan EDM, Rahbé Y, Duport G, Jouanin L, Kaiser L, Giordanengo P (2005) Effects of plant protease inhibitors, oryzacystatin I and soybean Bowman-Birk inhibitor, on the aphid *Macrosiphum euphorbiae* (Homoptera, Aphididae) and its parasitoid *Aphelinus abdominalis* (Hymenoptera, Aphelinidae). J Insect Physiol 51(1):75–86
99. Carrillo L, Martinez M, Alvarez-Alfageme F, Castañera P, Smagghe G, Diaz I, Ortego F (2011) A barley cysteine-proteinase inhibitor reduces the performance of two aphid species in artificial diets and transgenic Arabidopsis plants. Transgenic Res 20(2):305–319
100. Michaud D, Cantin L, Vrain T (1995) Carboxy-terminal truncation of oryzacysatin-II by oryzacytatin-insensitive insect digestive proteinases. Arch Biochem Biophys 322:469–474
101. Zhu-Salzman K, Koiwa H, Salzman RA, Shade RE, Ahn J-E (2003) Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. Insect Mol Biol 12(2):135–145
102. De Leo F, Bonade-Bottino M, Ceci L, Gallerani R, Jouanin L (1998) Opposite effects on *Spodoptera littoralis* larvae of high expression level of a trypsin proteinase inhibitor in transgenic plants. Plant Physiol 118(3):997–1004
103. Cloutier C, Jean C, Fournier M, Yelle S, Michaud D (2000) Adult Colorado potato beetles, *Leptinotarsa decemlineata* compensate for nutritional stress on oryzacystatin I-transgenic potato plants by hypertrophic behavior and over-production of insensitive proteases. Arch Insect Biochem Physiol 44(2):69–81
104. Mazumdar-Leighton S, Broadway RM (2001) Identification of six chymotrypsin cDNAs from larval midguts of *Helicoverpa zea* and *Agrotis ipsilon* feeding on the soybean (Kunitz) trypsin inhibitor. Insect Biochem Mol Biol 31(6–7):633–644
105. Strickland JA, Orr GL, Walsh TA (1995) Inhibition of *Diabrotica* larval growth by patatin, the lipid acyl hydrolase from potato tubers. Plant Physiol 109(2):667–674
106. Rispe C, Kutsukake M, Doublet V, Hudaverdian S, Legeai F, Simon J-C, Tagu D, Fukatsu T (2008) Large gene family expansion and variable selective pressures for cathepsin B in aphids. Mol Biol Evol 25(1):5–17

107. Hansen AK, Moran NA (2014) The impact of microbial symbionts on host plant utilization by herbivorous insects. Mol Ecol 23(6):1473–1496
108. Gil R, Latorre A, Moya A (2004) Bacterial endosymbionts of insects: insights from comparative genomics. Environ Microbiol 6(11):1109–1122
109. Chaudhary R, Atamian HS, Shen Z, Briggs SP, Kaloshian I (2014) GroEL from the endosymbiont *Buchnera aphidicola* betrays the aphid by triggering plant defense. Proc Natl Acad Sci U S A 111(24):8919–8924
110. Pinheiro P, Bereman MS, Burd J, Pals M, Armstrong S, Howe KJ, Thannhauser TW, Maccoss MJ, Gray SM, Cilia M (2014) Evidence of the biochemical basis of host virulence in the greenbug aphid, *Schizaphis graminum* (Homoptera: Aphididae). J Proteome Res 13(4):2094–2108
111. Tsuchida T, Koga R, Fukatsu T (2004) Host plant specialization governed by facultative symbiont. Science 303(5666):1989
112. Leonardo TE (2004) Removal of a specialization-associated symbiont does not affect aphid fitness. Ecol Lett 7(6):461–468
113. Ferrari J, Scarborough CL, Godfray HCJ (2007) Genetic variation in the effect of a facultative symbiont on host-plant use by pea aphids. Oecologia 153(2):323–329
114. McLean AHC, van Asch M, Ferrari J, Godfray HCJ (2011) Effects of bacterial secondary symbionts on host plant use in pea aphids. Proc Biol Sci 278(1706):760–766
115. Ferrari J, West JA, Via S, Godfray HCJ (2012) Population genetic structure and secondary symbionts in host-associated populations of the pea aphid complex. Evolution 66(2):375–390
116. Grenier A-M, Nardon C, Rahbé Y (1994) Observations on the micro-organisms occurring in the gut of the pea aphid *Acyrthosiphon pisum*. Entomol Exp Appl 70(1):91–96
117. Stavrinides J, McCloskey JK, Ochman H (2009) Pea aphid as both host and vector for the phytopathogenic bacterium *Pseudomonas syringae*. Appl Environ Microbiol 75(7):2230–2235
118. Leroy PD, Sabri A, Heuskin S, Thonart P, Lognay G, Verheggen FJ, Francis F, Brostaux Y, Felton GW, Haubruge E (2011) Microorganisms from aphid honeydew attract and enhance the efficacy of natural enemies. Nat Commun 2:348
119. Bansal R, Mian MAR, Michel AP (2014) Microbiome diversity of *Aphis glycines* with extensive superinfection in native and invasive populations. Environ Microbiol Rep 6(1):57–69
120. Pilon FM, Visôtto LE, Guedes RNC, Oliveira MGA (2013) Proteolytic activity of gut bacteria isolated from the velvet bean caterpillar *Anticarsia gemmatalis*. J Comp Physiol B 183(6):735–747
121. Broderick NA, Raffa KF, Goodman RM, Handelsman J (2004) Census of the bacterial community of the gypsy moth larval midgut by using culturing and culture-independent methods. Appl Environ Microbiol 70(1):293–300
122. Ferrater JB, de Jong PW, Dicke M, Chen YH, Horgan FG (2013) Symbiont-mediated adaptation by planthoppers and leafhoppers to resistant rice varieties. Arthropod Plant Interact 7(6):591–605
123. Hermisson J (2009) Who believes in whole-genome scans for selection? Heredity (Edinb) 103(4):283–284
124. De Mita S, Thuillet A-C, Gay L, Ahmadi N, Manel S, Ronfort J, Vigouroux Y (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. Mol Ecol 22(5):1383–1399
125. de Villemereuil P, Frichot É, Bazin É, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? Mol Ecol 23(8):2006–2019
126. Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. Philos Trans R Soc Lond B Biol Sci 367(1587):332–342
127. Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL (2010) Widespread genomic divergence during sympatric speciation. Proc Natl Acad Sci U S A 107(21):9724–9729
128. Chougule NP, Bonning BC (2012) Toxins for transgenic resistance to hemipteran pests. Toxins (Basel) 4(6):405–429

Chapter 5
Integrative Genomic Approaches to Studying Epigenetic Mechanisms of Phenotypic Plasticity in the Aphid

Mary Grantham, Jennifer A. Brisson, Denis Tagu, and Gael Le Trionnaire

Abstract Phenotypic plasticity is the nongenic variation in phenotype due to environmental factors. It is a common phenomenon in the animal kingdom that is not well understood at the molecular level. A tenable form of phenotypic plasticity for molecular research is polyphenism, which is an extreme form of phenotypic plasticity that results in discrete, alternative morphs. Epigenetic mechanisms have been hypothesized as the molecular regulators of polyphenism, in particular DNA methylation and chromatin remodeling. The pea aphid exhibits multiple polyphenisms including winged and wingless females during summer (wing polyphenism) and asexual and sexual morphs during summer and fall, respectively (reproductive polyphenism). The aphid is ideally situated for research into the molecular basis of polyphenism, with a sequenced genome and multiple transcriptomic studies that have begun identifying key molecular regulators of these two polyphenisms. The aphid also possesses the genes necessary for DNA methylation and chromatin remodeling. The pea aphid system is thus primed for future research into the epigenetic regulation of polyphenisms.

Abbreviations

ANT	Adenine nucleotide translocase gene
CHiP-seq	Chromatin immunoprecipitation sequencing
CRISPR	Clustered regularly interspaced short palindromic repeat
crRNA	CRISPR transcript

M. Grantham • J.A. Brisson
Department of Biology, University of Rochester, Rochester, NY 14627-0211, USA
e-mail: mary.chaffee@rochester.edu; jennifer.brisson@rochester.edu

D. Tagu • G. Le Trionnaire (✉)
UMR 1349 (INRA – Agrocampuus Ouest – University of Rennes I) IGEPP – Institute of Genetics Environment and Plant Protection, Domaine de la Motte, Rennes, Le Rheu cedex BP35327, 35657, 35653, France
e-mail: Denis.Tagu@rennes.inra.fr; gael.letrionnaire@rennes.inra.fr

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*,
Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_5

DNase-seq	DNAse I hypersensitive site mapping
DNMT	DNA methyltransferase
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
GO	Gene Ontology
H3K4me1	Mono-methylated lysine residue of histone H3
H3K4me2	Di-methylated lysine residue of histone H3
JH	Juvenile hormone
LSD1	Lysine-specific demethylase 1
MethylC-seq	Whole-genome bisulfite sequencing
NBAD	N-β alanyl dopamine
TALE	Transcription activator-like effector
TOL	Takeout-like gene

5.1 Introduction

Phenotypic plasticity is the nongenic variation in phenotype due to environmental factors [1]. Phenotypic plasticity is a common trait in both plants and animals that allows organisms or populations to adapt to changing local environments [1, 2]. Phenotypic plasticity can exhibit a gradient from subtle variation in the phenotype to extreme, discrete, morphological development. The extreme, discrete form of phenotypic plasticity is termed polyphenism. In polyphenism, two or more discrete morphs are produced in response to environmental stimuli.

How polyphenisms are established at the molecular level is poorly understood and thus an area of active research [3, 4]. The two primary molecular mechanisms that have been proposed for regulation of polyphenism are epigenetic and neuroendocrine [5]. Epigenetic mechanisms can alter the transcription of genes by changing the accessibility of DNA to binding factors such as transcription factors, silencers, or enhancers [6]. The nervous system and endocrine system often work together; the communication between these two systems is the neuroendocrine system [7]. Neurotransmitters, neuropeptides, and hormones that form the neuroendocrine system are likely involved in polyphenic induction [8, 9]. The nervous system sends signals throughout the body using neurons that are triggered by neurotransmitters and neuropeptides [7]. The endocrine system modulates hormone levels that are directly secreted into the circulatory system and can be transmitted quickly across the body [10]. Epigenetic and neuroendocrine mechanisms are not mutually exclusive, and evidence from recent studies indicates that they may function in different portions of the same pathway to establish alternative morph development.

Polyphenism is a widespread trait among insects. One of the best-studied examples of polyphenism is the caste system of eusocial insects such as the honeybee, *Apis mellifera*, and carpenter ants, *Camponotus floridanus*. The honeybee exhibits a caste polyphenism between female workers and queens. Larvae that are fed on a normal diet develop into workers whereas larvae that eat royal jelly develop into

queens. Epigenetic mechanisms control this switch, at least in part. When the enzyme responsible for de novo DNA methylation is knocked down by RNA interference (RNAi), larvae that should develop as workers develop instead as queen-like bees [11]. At the genome level, DNA methylation patterns are significantly different between queens and workers and influence the insulin-signaling pathway [12, 13]. Further, the active ingredient of royal jelly is a histone posttranslational modification inhibitor [14]. Posttranslational modifications of histone tails alter the accessibility of DNA to transcription factors. These data suggest that caste regulation in honeybees involves a complex interaction between changes in specific epigenetic marks such as DNA methylation and posttranslational modifications of histone proteins. Ants also exhibit an environmentally induced caste system that can be distinguished by epigenetic marks. Simola et al. [15] identified H3K27 acetylation as a modification that differed between the major and minor worker castes in the carpenter ant. In addition, differentially methylated genes have been identified between castes in both *C. floridanus* and *Harpegnathos saltator* ant species [15].

Neuroendocrine regulation of polyphenisms has been a primary research avenue due to its potential to control multiple downstream phenotypes in a coordinated fashion. Further, hormones are often responsive to environmental cues [10]. Multiple hormones have been identified as regulatory features of polyphenic induction. For example, ecdysteroids modulate butterfly wing patterns [16], the peptide hormone insulin modulates honeybee caste determination [17] and dung beetle horn size dimorphism [18], and serotonin, a neuropeptide, modulates locust gregarious versus solitarious phenotypes [19]. Endocrine mechanisms that mediate polyphenisms have been discussed elsewhere (see [10, 20]). While this field of research has offered many insights into the regulation of polyphenisms such as the examples listed above, the realization that the endocrine system likely works in concert with epigenetics has begun to resolve some of the limitations of purely endocrine-focused research. For example, in the honeybee, the knockdown of insulin receptors in queen-destined embryos leads to worker morphology, but does not fully switch the morph. This results in an intermediate phenotype with worker pollen baskets and queen ovarioles [17]. It is possible that epigenetic or other mechanisms work in concert with these hormonal signals to regulate finer features of morph development.

While both epigenetic and endocrine mechanisms are important factors to consider with respect to the mechanistic basis of polyphenism, this review will focus primarily on epigenetics in alternative morph induction and will bring in studies of endocrine function specifically where they interact with epigenetics. Below, we will introduce the pea aphid, *Acyrthosiphon pisum*, the model organism in aphid research, and present what is known about epigenetic mechanisms in this species. We will evaluate molecular tools that can be used to analyze gene expression and genomic editing; these have potential for functional studies in the aphid. Lastly, we will discuss the role of epigenetics in the establishment of alternative phenotypes in the aphid and propose future directions for epigenetic research.

5.2 A Model System for Phenotypic Plasticity Research: The Pea Aphid

The aphid system has benefited from a long history of research. Various species of aphids can be devastating crop pests, which has resulted in decades of research on how aphids reproduce and how they affect crop plants, as well as their general ecology. Further, aphids have a fascinating, complex life cycle that includes alterations between various phenotypes through polyphenism. This makes them an ideal system for studying phenotypic plasticity. We will focus on two of these polyphenisms, the wing polyphenism and the reproductive polyphenism.

5.2.1 Ecology and Life History

Aphids are soft-bodied insects of the order Hemiptera. Among approximately 4400 species of aphids, over 250 are agricultural or horticultural crop pests [21] which are prevalent in temperate regions, causing extensive damage to agriculture every year [22]. The estimated annual loss of crops due to aphids in the United States of America is in the hundreds of millions of dollars [23, 24]. Aphids feed on plant phloem by inserting their mouthparts between the cells extending to the sieve tubes, where they pierce the cell and begin feeding. In addition to feeding damage, aphids are highly effective plant virus vectors that can cause significant damage to crops [22]. Aphid species can range from highly specific host plant pests, such as soybean aphids, *Aphis glycines*, which feed exclusively on soybean species, to generalists like pea aphids which exploit up to 30 different legume species as hosts.

Aphids are excellent model organisms for studying polyphenism. Aphids express multiple polyphenisms with many alternative morphs, the two best-studied being wing and reproductive polyphenism. Briefly, winged and wingless morphs are produced during summer conditions in response to environmental stress (the wing polyphenism). Sexual females are produced from asexual females in response to photoperiod shortening and secondarily to decreasing temperature (the reproductive polyphenism). Aphids are easy to maintain in the laboratory, and both polyphenisms are inducible under controlled laboratory conditions.

These polyphenisms can be best explained in detail within the context of the pea aphid's life cycle (Fig. 5.1). During the summer months, viviparous (live-bearing), asexual females reproduce by parthenogenesis, and no males are found in the population. Because the mode of parthenogenesis they use lacks chromosomal recombination, mothers and their daughters are genetically identical [25]. The summer morphs can be winged or wingless depending on environmental conditions (explained more below). In autumn, sexual individuals are produced. Sexual morphs include oviparous (egg-bearing) females and males. Sexual females are genetically identical to asexual females, while males differ from females in that they have only one X chromosome (XO) instead of two (XX). Males may be winged or wingless,

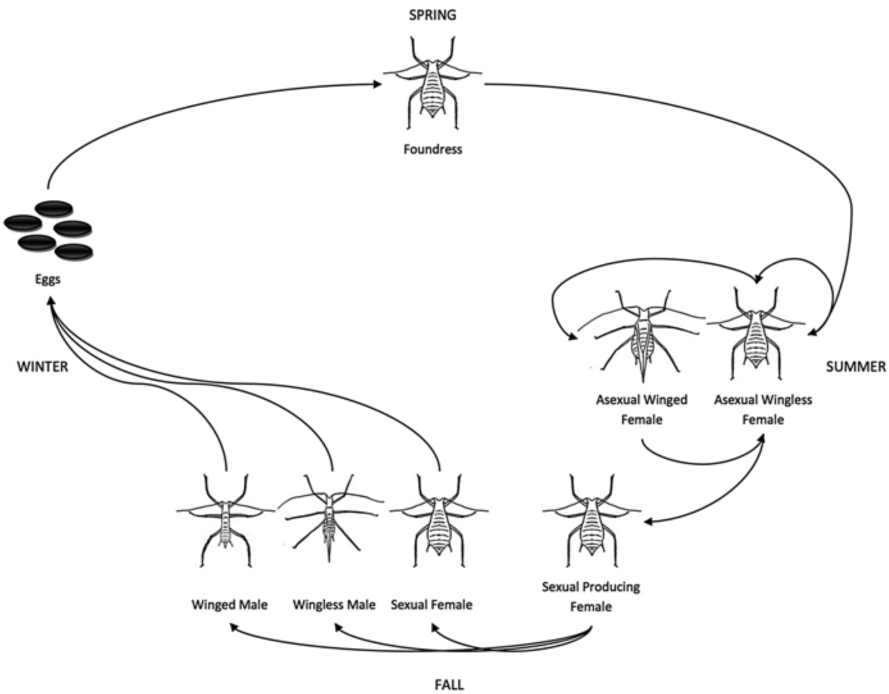


Fig. 5.1 Pea aphid life cycle in spring; the foundress emerges from an egg and viviparously produces asexual females through parthenogenesis. The asexual females will continue producing asexual females (winged or wingless) throughout the summer months. As the photoperiod shortens and the temperature drops, the asexual females will produce sexual-producing females. The sexual producing females will viviparously produce oviparous females and males. The fertilized eggs from the oviparous females and males will overwinter and a new foundress will emerge in spring

although wingedness in males is genetically determined by the single X-linked locus, *aphicarus* [26, 27]. Males and sexual females mate and females produce fertilized eggs which overwinter via diapause. In spring, a foundress emerges from the egg as a viviparous, asexual female that restarts the life cycle.

5.2.2 Polyphenisms and Alternative Morph Characteristics of the Pea Aphid

As noted above, an adult asexual female can produce winged or wingless offspring in response to environmental conditions. This is the wing polyphenism. The asexual females can sense suboptimal conditions such as lack of food or the presence of predators and pass this environmental information to the embryos developing within their ovaries (reviewed in [28]). Once born, these offspring then proceed through development (four nymphal instars) and mature with a winged phenotype which

allows them to migrate to more optimal conditions. In contrast, offspring developing in optimal conditions will be wingless. Wingless females are able to devote the bulk of their energy into the production of offspring rather than flight, optimizing reproduction in the population. The production of winged versus wingless females allows for a trade-off between dispersal by winged adults from poor conditions and a high reproduction rate of wingless adults while in good conditions, as in many other insects that exhibit winged and wingless phenotypes (reviewed in [29]).

The wing polyphenism of the pea aphid is comprised of whole-body changes. Winged and wingless females differ in more than just whether they have wings. The winged morph possesses functional flight muscles and heavier sclerotization of the cuticle on the head and thorax [30]. Winged females also exhibit an expanded sensory system with ocelli on the vertex of their heads, longer antennae, and more secondary rhinaria (sense organs) on the antennae [31–33]. The behavior of the winged morph is agitated whereas the wingless morph exhibits a sedentary behavior [30]. In contrast, the wingless morph has higher fecundity and matures faster [30].

The second polyphenism we will consider here is the reproductive polyphenism. This polyphenism occurs when asexual females (called sexuparae) perceive the photoperiod decrease in autumn. This photoperiod shortening is translated into a signal that is then transduced and transmitted to developing embryos. The photoperiod signal can also be buffered by temperature. These embryos will then grow up to be oviparous females or males. The signal mainly affects the germline cells of the embryo. This causes a switch from a diploid nonrecombining germline (in asexual embryos) to a true haploid germline (in sexual embryos). Thus, the biggest difference between asexual females and sexual females is that the former have diploid embryos developing in their ovaries while the latter have eggs in their ovaries. Once born, in addition to having true sexual gonads that will produce haploid gametes (eggs in females or sperm in males), sexual individuals display strong morphological differences compared to asexual morphs. For example, sexual females have different leg and antennae morphology compared to asexual females. Males are significantly smaller than females, have male sexual characters such as testes and a visibly different genital plate, and have more secondary rhinaria than any of the female morphs [32]. Males tend to be less active in terms of feeding and more active in terms of general movement, probably due to investing more energy in their reproductive behavior.

5.2.3 Existence of Epigenetic Machinery in the Pea Aphid

Epigenetic mechanisms are gaining traction as a possible molecular regulator of polyphenic traits in insects [3, 4, 34]. Epigenetics is the study of heritable changes in genome function that occur without alterations of the DNA sequence [35]. By not altering the DNA sequence, changes by epigenetic mechanisms are plastic and are

inherently reversible [35]. Among the best-studied epigenetic mechanisms for altering an organism's gene expression are DNA methylation and chromatin remodeling [6, 36], both of which can contribute to changes in transcription, including up- and downregulation of gene expression and alternative splicing. The pea aphid possesses the necessary enzymes for epigenetic regulation by DNA methylation and chromatin remodeling.

DNA methylation is the addition of a methyl group to genomic cytosines. Methylated DNA can reduce binding efficiency of DNA-binding proteins though steric hindrance or attract specific DNA-binding proteins [37–39]. Cytosines are methylated by a family of enzymes called DNA methyltransferases (DNMTs). DNA methylation is maintained through DNA replication by DNMT1s, and de novo DNA methylation is achieved by DNMT3s [35]. The pea aphid possesses both DNMTs necessary for an active DNA methylation system, including two copies of DNMT1 and one copy of DNMT3 [40].

The location of methylated cytosines in the genome results in different transcription levels. Gene body methylation is associated with actively transcribed genes, and promoter methylation is associated with gene silencing [37, 41]. Additionally, gene body methylation is likely involved in the regulation of alternative splicing [42]. Interestingly, insects do not methylate their DNA at promoters; they primarily display gene body methylation [34, 43]. However, some insects display methylation of transposable elements [44]. Consistent with this trend of insect DNA methylation, pea aphids nearly exclusively methylate gene bodies, which are correlated with actively transcribed genes [40]. Further, like other insects, pea aphids exhibit a low overall level of cytosine methylation in their genomes; only about 0.69 % of cytosines are methylated [40]. And finally, DNA methylation shows a differential pattern in gene bodies among the female winged, wingless, and sexual morphs (Brisson, Tagu, and Edwards, *unpublished*).

Chromatin remodeling is the mechanism by which chromatin structure is altered, affecting DNA accessibility. DNA is tightly wrapped around histone proteins to form nucleosomes (reviewed in [6]). The highly organized compaction of DNA interferes with most reactions that require DNA binding, which include transcription, DNA replication, and DNA repair [6]. Chromatin remodeling creates "open" regions in the DNA that are nucleosome depleted and allows for the binding of transcription factors, activators, repressors, and other DNA-binding elements [45]. This remodeling is achieved by histone modifications (such as methylation, phosphorylation, acetylation, ubiquitination, and ADP-ribosylation [46]) and ATP-dependant chromatin remodeling complexes [47, 48].

Pea aphids have a diverse suite of histone-modifying enzymes that form chromatin remodeling complexes such as SWI/SNF, CHD1, ISWI, and NURD genes [49]. Further, immunofluorescence studies of pea aphid chromatin have detected several histone modifications such as the methylation of lysines 4 and 9 of histone H3 [49–51].

5.2.4 How Do We Study the Effect of DNA Methylation or Chromatin Structure on Genes in a Species Like the Pea Aphid?

The mechanisms necessary for epigenetic regulation have been identified in the pea aphid, but how is it possible to identify epigenetic regulation involved in the control of polyphenism? As noted above, the traits associated with the alternative phenotypes in pea aphid polyphenisms occur throughout the whole organism. This is especially evident in the wing polyphenism where sensory organs, cuticle, wing formation, musculature, and behavior are all modulated to produce the alternative morphs. Thus, we expect epigenetic regulation to affect genes at the genome-wide level, so the best way to begin studying it will be a genome-wide approach.

Resolution at the nucleotide level of DNA methylation can be accomplished by bisulfite conversion of the DNA. Using this method, unmethylated cytosines are converted to uracils and methylated cytosines are unchanged. The converted sequences are then compared to the untreated sequence to identify methylated cytosines. This can be done on a genome-wide scale with whole-genome bisulfite sequencing (MethylC-seq) [52] or using a targeted-gene approach with PCR and Sanger sequencing [53]. Either the targeted-gene or genome-wide approach can be used to analyze whole bodies of alternate polyphenic morphs for overall correlation with DNA methylation patterns or in a tissue-specific manner to identify trait-specific DNA methylation patterns.

Identifying regions of open chromatin on a genome-wide level can be done effectively by two different techniques: formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) and DNAse I hypersensitive site mapping (DNase-seq) [45]. Open chromatin regions identified by FAIRE-seq and DNase-seq are usually binding sites for transcription factors, enhancers, silencers, and insulator proteins [45]. FAIRE is a highly sensitive method that isolates open chromatin regions by utilizing formaldehyde cross-linking between the DNA and bound proteins that are removed during phenol/chloroform DNA isolation. This method can be used to detect open chromatin regions even if they are present in only a small number of cells in a sample [54]. DNase-seq uses an endonuclease to fragment the DNA and size selection to identify regulatory regions. When FAIRE-seq is used in concert with DNase-seq, they cross validate many regions such as active transcription start sites [45]. In addition, FAIRE-only or DNase-only enriched sites identify distinct genomic regions [45]. DNase-only sites are found to be within 2 kb of transcription start sites and within 5′ exons and introns [45], and FAIRE-only sites tend to be in non-promoter intergenic regions and internal exons and introns [45]. Additionally, FAIRE-seq and DNase-seq are associated with different histone modifications [45]. A combination of both methods would be a valuable, unbiased approach to scan the genome for differentially open regions associated with the establishment of alternative phenotypes.

The pea aphid is primed for these kinds of genome-wide approaches because the genome was recently sequenced [55]. The genome assembly is available at AphidBase accompanied with multiple tools such as BLAST, Gene Ontology (GO) annotations, and a dedicated Galaxy Server for additional genomic tools [56].

Once the data are generated, genome-wide epigenetic maps can be integrated with transcription data obtained by RNA-seq or microarray experiments for identification of candidate mechanisms and master gene regulators of polyphenisms. Once candidate mechanisms and master regulator genes have been identified, how do we further test their specific role in polyphenic regulation? Some epigenetic mechanisms can be knocked down by chemical inhibition. For example, the DNMTs responsible for DNA methylation can be inhibited by multiple chemical interactions. One of the most promising DNMT inhibitors is zebularine, which, as shown in mouse studies, effectively knocks down both DNMT1 and DNMT3 by forming a covalent bond with them [57–59]. Zebularine has been used in the pea aphid by injection, affecting growth rate and fertility [60]. Chemical inhibition is thus a promising method for inhibiting these enzymatic processes. RNAi can also be used to knock down the transcript level of particular DNMTs. RNAi can be introduced into aphids by injection for the knockdown of gene expression in specific tissues (primarily abdominal tissues) or by incorporation into artificial media (primarily gut and salivary tissues) [61, 62]. However, RNAi has been generally ineffective for whole-body gene knockdown in aphids because dsRNA is rapidly degraded in the hemolymph [63]. A limitation of both chemical and RNAi knockdown of DNMTs using these methods is that they affect all DNA methylation across the genome. Approaches that target specific DNA methylation or histone modification at particular genes are often needed to identify the functional significance of particular modifications.

Clustered regularly interspaced short palindromic repeats (CRISPR) and transcription activator-like effectors (TALEs) are exciting new methods of genome editing that can used to target specific places in the genome. The CRISPR system was identified from bacteria and archaea as an immune mechanism to protect against invading viruses and plasmids [64–66]. The CRISPR transcripts (crRNA) function as guide RNAs for Cas9, an endonuclease that causes dsDNA breaks (reviewed in [67]). This system is programmable in multiple ways to create dynamic, sequence-specific genomic editing tools that can be used to target virtually any locus [68]. By mutating the Cas9 protein, different outcomes can be achieved such as dsDNA or ssDNA cleavage to create gene knockouts, inactivation of active sites on Cas9 to achieve gene knockdown without DNA cleavage, and targeted fusion of transcription factors to modulate expression (activation or repression) of specific genes [69]. To date, there is no published report of using the CRISPR/Cas9 system for modulation of the epigenetic state of a gene. However, this approach/strategy has been discussed as the next step in CRISPR/Cas9 engineering by fusing a histone-modifying protein or DNA demethylase to the Cas9 protein [70, 71]. The CRISPR/Cas9 system has been shown to be effective in many eukaryotic species, making it an exciting new tool to study gene function [68, 72, 73].

TALEs were discovered as proteins secreted by the plant pathogen *Xanthomonas*. TALEs generally consist of a specific DNA-binding domain and nuclease that cause dsDNA breaks resulting in selective gene knockouts [74]. The nuclease can be swapped out with other proteins, creating some of the most exciting uses of TALEs. Of particular interest is the ability to manipulate the epigenetic state of a gene through this swapping out process [75, 76]. TALEs allow the fusion of enzymes like histone demethylases or DNA demethylases to alter the epigenetic state of a specific sequence [75, 76]. For example, Mendenhall et al. (2013) [76] fused a lysine-specific demethylase 1 (LSD1) to a TALE to test its effectiveness for epigenetic state modulation. LSD1 specifically demethylates mono- and di-methylated lysine residues of histones (H3K4me1 and H3K4me2) [77]. Both of these histone modifications are associated with enhancer sites for active transcription [78]. The TALE–LSD1 effectively knocked down specific gene expression, and H3K4 was demethylated up to threefold [76].

A summary of the different functional tools that can be used to study epigenetic modifications is listed in Table 5.1. RNAi has been used before in the pea aphid, as noted above, as has zebularine [60]. TALEs are currently being tested in the pea aphid and CRISPR is likely to be tested in the near future. The need for these kinds of resources will likely drive innovations that can eventually be applied to the pea aphid.

Table 5.1 Genomic editing tools for studying gene expression and epigenetic mechanisms

Mechanism	Function	Activity	Specificity	References
RNAi	Gene expression knockdown	Degradation of transcripts	Sequence specific and programmable	[11, 61, 63]
Zebularine, 5-azacytidine, 5-aza-2′-deoxycytidine	DNMT inhibition	Binds DNMT active site	Limited to DNMTs	[57–60, 79]
CRISPR	Gene knockout	dsDNA Cleavage	Sequence specific and programmable	[64, 68, 69]
	Gene knockdown	Inactive Cas9 (CRISPRi)	Sequence specific and programmable	[80]
TALEs	Epigenetic modulation	DNA demethylase, histone modification enzyme	Sequence specific and programmable	[74–76]

DNMT DNA methyltransferase, *CRISPR* clustered regularly interspaced short palindromic repeat, *TALEs* transcription activator-like effectors, *CRISPRi* Inactive Cas 9

5.3 Morphological Maintenance in Adults

We have discussed the tools available for examining epigenetic modifications. Given these technologies, the focus now turns on which stages and which tissues to apply these approaches. Only by careful consideration of the alternative phenotypes and their development can we anticipate the best experimental strategies for beginning to understand the epigenetic basis of polyphenism. Here, we consider two different but complementary concepts with respect to polyphenisms. The first is morph *maintenance*. Morph maintenance refers to the molecular mechanisms that underlie the development and functioning of the different phenotypes after their developmental trajectories have been decided. These are processes far downstream of morph *determination*, which is the second concept that we consider below.

The adult alternative morphs are highly divergent from each other phenotypically and consist of multiple morphological differences. Maintaining these different morphs requires a suite of expressed genes specific to the morph. A large number of gene expression differences have been recognized in the alternative adult phenotypes.

Ghanim et al. (2006) [81] were the first to analyze gene expression differences between winged and wingless adults using the green peach aphid (*Myzus persicae*). They identified 31 differentially expressed genes. Three genes were chosen for further analysis: adenine nucleotide translocase (*ANT*), takeout-like (*TOL*), and *OS-D*. All three genes showed significantly increased expression in winged females. The three genes were also tested for spatial expression differences. As expected from *Drosophila*, *OS-D* was primarily expressed in the antennae, with the second highest expression level in legs. *OS-D* has been implicated as an odorant-binding protein with a chemosensory role, but the evidence for this function is weak. *ANT* is involved in ADP/ATP transport and was significantly expressed in the thorax where flight muscles and wings are located. Lastly, *TOL* has been identified in *Drosophila* as participating in circadian rhythm and food status. In the green peach aphid, *TOL* was most highly expressed in the head, followed closely by abdomen. Brisson et al. [82] compared winged to wingless morphs in the pea aphid by microarray analysis. The transcripts with significantly differential expression consisted of genes associated with energy production and flight muscle maintenance [82].

RNA-seq has been performed on five pea aphid adult morphs: asexual females (winged and wingless), sexual females, and males (winged and wingless) [83]. Large transcriptional differences were found between the different sexual morphs (asexual, sexual, and male). GO analysis of the differentially expressed genes showed enrichment of gene groups specific to each morph such as cell cycle and chromatin in sexual females, development and cellular differentiation in asexual females, and ion channel, hormone, and hydrolase activity in males [83]. Whole-body samples were used for this study, including embryos or eggs for females, and most likely the GO terms associated with the females are related in part to their carrying developing offspring.

Liu et al. [84] compared the transcriptomes from whole bodies of sexual and asexual females in the cotton aphid, *Aphis gossypii*, and found many transcriptional differences between the two morphs (1614 genes with higher expression and 2238 with lower expression in sexual females relative to asexual females). Multiple genes

related to photoperiod detection were differentially regulated between sexual and asexual females. Additionally, 19 cuticle-related genes were differentially expressed in the cotton aphid females. Cuticle-related genes have also been found to differ between sexual and asexual females in the pea aphid [84, 85]. Cortes et al. [85, 86] examined differentially expressed genes between pea aphids reared under long- or short-day conditions. Cuticle-related transcripts were differentially regulated, in addition to genes related to gene regulation such as mRNA processing and chromatin remodeling-related genes.

Useful information has been garnered about the maintenance of the different alternative morphs from these studies; however, they offer little insight into the induction and development of the alternative morphs. By adulthood, the patterns of genome expression have been irreversibly altered, and the development of morph-specific traits has ceased. To understand how genome expression is radically altered in response to changing environmental conditions, we need to learn how the signal is received by the adult, interpreted, and transmitted to the developing offspring. Furthermore, how the signal is integrated by the offspring and used to modulate genome expression will require studies of embryonic gene expression and regulation through the inducible period of development. These processes occur during morph determination.

5.4 Morph Determination

Our working hypothesis for alternative morph induction in both polyphenisms is that the female receives a morph-inducing signal from the environment. The signal is translated and amplified through a signal transduction cascade that travels to the offspring. The development of the offspring is then altered in response. Based on this model, which is illustrated in Fig. 5.2, it is important to study the signal cascade starting from the mother and follow it through to the offspring.

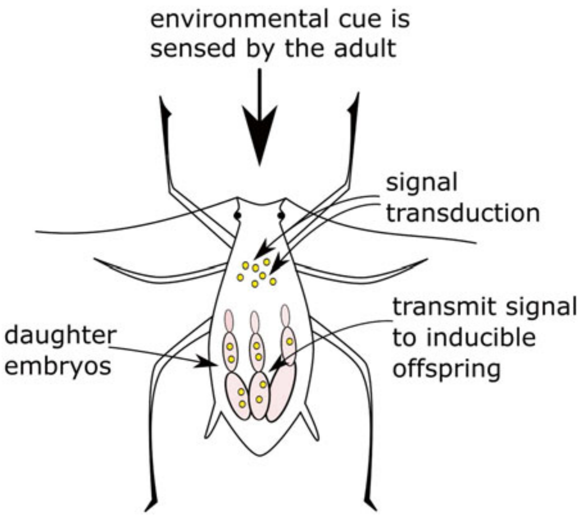


Fig. 5.2 Hypothesized transduction of environmental signal to offspring. The adult female perceives the environmental signal through her sensory system. The signal is synthesized to a cue that can be transmitted to offspring. The offspring receive the signal and alter their developmental trajectory

5.4.1 Sexual Polyphenism

The establishment of the reproductive polyphenism can be divided into two parts: (1) the initial steps of perception and transduction of the photoperiodic signal that occur within the heads of parthenogenetic individuals (also called sexuparae) and (2) the asexual to sexual switch during embryogenesis as a direct response of the transduced decreasing photoperiodic signal.

5.4.1.1 Genetic Programs Involved in the Initial Steps of Photoperiodic Signal Perception and Transduction

Physiological studies in the 1980s suggested that the photoperiodic response involves neuroendocrine control that occurs within the head of aphids. In particular, one group of neurosecretory cells from the protocerebron might be involved in the transduction of the signal since the micro-cauterization of those cells abolishes the response [87]. The recent development of genomic tools in the pea aphid allowed the comparison of head transcriptomes from short- and long-day reared aphids and led to the identification of putative key regulatory genes in this process [86, 88, 89]. Microarray analyses revealed overrepresentation of some functional categories among the differentially expressed transcripts between the two conditions. Several transcript coding for proteins involved in nervous system function and endocrine signaling, such as the insulin signaling pathway, were differentially regulated, confirming earlier hypotheses (summarized in Fig. 5.3). Interestingly, calnexin and arrestin transcripts, known in *Drosophila* to be involved in photoactivation of the photopigment rhodopsin, were also differentially expressed, suggesting a possible role for photoreception in the process. More surprisingly, a significant number of transcript coding for cuticular proteins containing a chitin-binding domain (RR domain) were downregulated under short-day conditions, suggesting a possible modification of cuticle structure in response to day length shortening. N-β alanyl dopamine (NBAD) is known to be a linking component of the chitin-cuticular protein network. Transcript coding for *ebony* and *black*, respectively, involved in β-alanine synthesis and conjugation, was also less abundant in short-day reared aphid heads, suggesting a possible desclerotization of cuticle structure.

A more recent study showed that *ddc* (dopa decarboxylase) and *th* (tyrosine hydroxylase) transcripts – coding for enzymes involved in the dopamine biosynthesis pathway – had identical profiles, suggesting a decreased synthesis of dopamine under short photoperiod [92]. Whether reduced dopamine synthesis has a consequence on cuticle structure (through NBAD synthesis) or on neuronal signaling is not known but is worth investigating. The dopamine pathway is involved in the solitarious to gregarious phase transition of *Locusta migratoria*, another well-known example of insect polyphenism [93]. These transcriptomic analyses thus allowed the identification of several candidate transcripts potentially involved in the

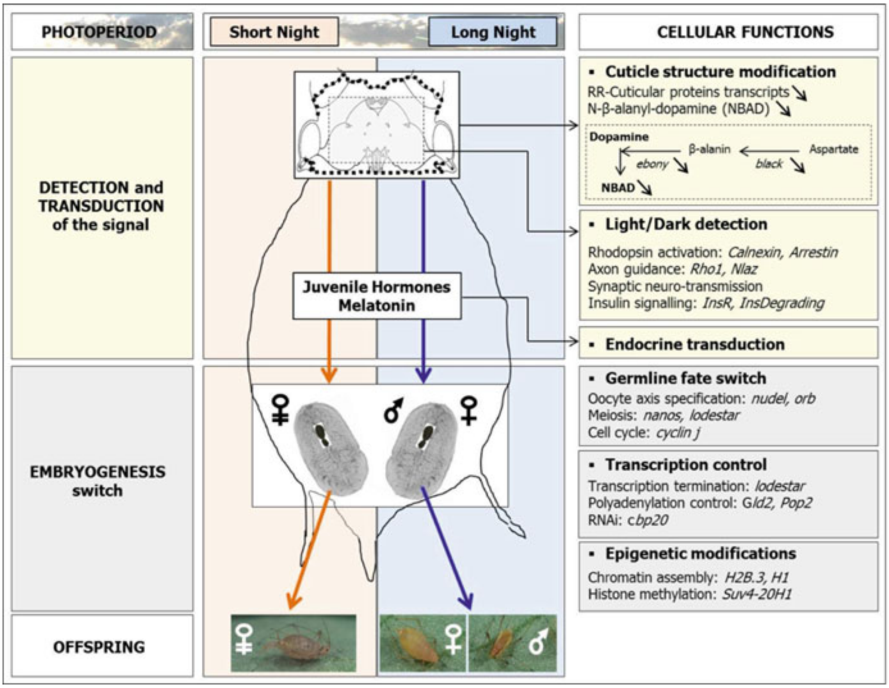


Fig. 5.3 Reproductive polyphenism regulation: genetic programs associated with the initial steps of photoperiodic signal transduction and later steps of sexual to asexual embryogenesis switch. This figure summarizes the results of two recent large-scale transcriptomic studies that, respectively, aimed at identifying: (1) the differentially expressed transcripts between the head of long-day reared (asexual reproduction) and short-day reared (sexual reproduction) aphids in order to investigate the initial steps of photoperiodic signal detection and transduction [86] and (2) the transcripts regulated between sexual and asexual embryos so as to identify the key genetic programs associated with the transition from a sexual to an asexual embryonic germline [90] (Reprinted with permission from Le Trionnaire et al. [91])

perception and transduction of the photoperiodic signal and confirmed earlier physiological studies suggesting that photoreception and neuroendocrine signaling are likely key signaling pathways involved in this phenomenon.

5.4.1.2 Genetic Programs Involved in the Asexual–Sexual Embryogenesis Switch

Juvenile hormone (JH), which is known to be involved in the regulation of key developmental pathways in insects, is also involved in regulation of the reproductive polyphenism in aphids. Hardie and Lees [94] performed ectopic applications of kinoprene – an analogue of JH – on sexuparae individuals (e.g., reared under short photoperiod to produce sexual offspring) and observed a reversion of the reproductive mode, with asexual aphids instead of sexual ones produced in the offspring. JH

is thus probably an important intermediate between the photoperiodic signal transduction step and the embryogenesis switch. Gallot et al. [90] reproduced this experiment and collected sexual and asexual embryos from nontreated and kinoprene-treated sexuparae individuals, respectively. A large-scale transcriptomic analysis was then performed using oligonucleotide microarrays to compare the gene expression profiles of both types of embryos. Such a fine-tuned and synchronized protocol allowed the identification of a limited set (33) of differentially expressed transcripts (summarized in Fig. 5.3). Among those genes, several were involved in oogenesis (*orb* and *nudel*), posttranscriptional regulation (*pop2*), cell cycle control (*cyclin J*), and epigenetic mechanisms (see below). In situ hybridization of those transcripts showed that most of them were specifically expressed in the germline cells of those embryos. Kinoprene treatment thus appears to have a direct consequence on embryos' germline fate (switching from a haploid to a diploid germline). JH might thus be at the very end of the signaling cascade, starting with photoperiod perception and finishing with embryonic germline fate switch.

Chromatin remodeling transcripts were also identified as being differentially expressed in this study [90]. The differential expression of transcript coding H1 and H2B.3 histone proteins as well as a histone methyltransferase (Suv4-20H1) between sexual and asexual embryos (combined with their specific localization within the germline cells) strongly suggests that epigenetic regulation might be a key player in the regulation of the reproductive polyphenism. We can make the hypothesis that neuroendocrine signaling in response to the decreasing photoperiod signal might target specific receptors within the embryo germline. As a consequence of this putative receptor activation, some chromatin remodeling events might occur within germline cells. This would allow the differential opening or modulation of the expression of specific genomic regions containing gene and regulatory elements (enhancers, silencers, or insulators) associated with the sexual/asexual germline fate. The control and the regulation of reproductive polyphenism in aphids might thus involve a combination of neuroendocrine and epigenetic mechanisms.

These studies have demonstrated the utility of focusing on signal perception by the mothers and corresponding changes in the developing embryos within the affected mothers. Future studies will focus on epigenetic profiling of these same stages to provide insight into how transcriptional changes are regulated by DNA methylation and/or histone modifications.

5.4.2 Wing Polyphenism

As previously stated, it is important to study the signal transduction cascade from the initial reception by the adult female to the developmental alterations that lead to alternative morph development in the embryos. Similar to the reproductive polyphenism, the wing polyphenism can also be separated into two stages: (1) the reception of the inducing signal (i.e., crowding, predation, plant stress) by the asexual adult female followed by signal transduction and transmittance to the developing

embryos and (2) the reception of the inducing signal in the embryos causing altered embryogenesis that ultimately results in alternative adult phenotypes (winged and wingless).

5.4.2.1 Genetic Programs Involved in Reception of the Wing-Inducing Signal and Further Signal Transduction and Transmittance

Unlike the reproductive polyphenism, the neuroendocrine basis of the wing polyphenism is still unknown. Historically, JH was of great experimental interest, in part because an adult wingless female resembles a juvenile adult winged female. Despite intensive investigations into the possible involvement of JH in suppressing wing development in aphids during the 1960s–1980s, no strong evidence supporting such a role for JH has been obtained (reviewed in [30, 95]). Therefore, all that is known about the genetic programs underlying reception and integration of wing-inducing signals comes from a limited number of studies examining transcript abundance.

The first study to examine this issue used vetch aphid, *Megoura viciae*, adult females that were crowded or not crowded [96]. Multiple differentially expressed transcripts were identified in maternal tissue using differential display. The most interesting findings were two ubiquitin genes (*ubiquitin ligase* and *ubiquitin-activating enzyme E1*) and two possible ubiquitin-associated genes (*wingless* and *naca*) that were all expressed at higher transcript abundance in winged offspring-producing (crowded) females [96]. Thus, protein degradation may be required to mediate the switch from producing wingless offspring to producing winged offspring in response to high-density conditions.

A subsequent, unpublished study (Vellichirammal et al., personal communication) examined transcriptional changes at the genome-wide level by RNA-seq in wingless females that were exposed to wing-inducing stimuli or no stimuli. 1509 transcripts were more abundant in winged offspring-producing females than wingless offspring-producing females, and 2006 transcripts were more abundant in wingless offspring-producing females than winged offspring-producing females. An analysis of these transcripts using GO terms revealed that both hormonal and epigenetic terms were enriched in wingless offspring-producing females. Specifically, expression of chromatin remodeling and ecdysone-associated genes was at lower levels in winged offspring-producing females. Ecdysone is a steroid hormone. Studies in *Drosophila* show that the expression of ecdysone in adults is dynamic and appears to be involved in multiple systems and tissues including tissue-specific gene repression (reviewed in [97]). Taken together, these results suggest that the wing polyphenism utilizes both epigenetic and hormonal regulation for fine-tuning of the induction process.

5.4.2.2 Genetic Programs of the Wingless to Winged Switch During Embryogenesis

Further investigations are needed to elucidate the pathway from signal reception in the adult female to modulation of development in the offspring. To date, no RNA-seq or microarray analysis has been published for morph-verified embryos. One serious obstacle in continuing research into the embryos has been the identification of the inducible stages during embryogenesis. Recently, the critical induction period for developmental plasticity in the wing polyphenism during embryogenesis was identified. During development, aphids undergo 20 embryonic stages [98], and stages 15–19 have been tentatively identified as the inducible period [99]. During stage 19, the embryos are locked into a developmental pathway; thus, this stage consists of a mix between inducible and determined embryos [99]. Two conflicting hypotheses have been proposed for embryos at stages 15–17: these embryos are inducible or the signal accumulates in the adult and does not induce the embryos until they are older [99]. So, stage 18 embryos can be used reliably for studying the embryonic induction period. Once transcriptomic data have been collected on stage 18 embryos, candidate genes for master regulators of the wing polyphenism can be identified.

Current studies on the wing polyphenism have emphasized the importance of transcriptomic data and the need for further research into induction of wing development. Future work on the wing polyphenism will focus on transcriptional information in the embryo and how the transcriptional changes are regulated by the epigenetic mechanisms of DNA methylation and chromatin remodeling.

5.5 Future Work

A significant amount of genomic data is now available for the aphid genome, ranging from RNA-seq data to DNA methylation patterns (transcriptome and epigenetic mechanisms). How do we interpret these data in terms of alternative morph induction? The current data have begun to clarify which mechanisms are the likely regulators of polyphenism in the pea aphid. However, to fully understand the mechanism, more information is needed. Specifically, we need more embryonic data on the wing polyphenism to obtain a more complete picture of how environmental signals are recognized by the mother and employed by the offspring to change their development. A significant challenge for studies in the wing polyphenism is the lack of a chemical stimulant. The sexual morph has the advantage of kinoprene studies to synchronize the production of asexual and sexual embryos. Further, as shown by Ishikawa et al. [96] and Vellichirammal et al. (personal communication), highly wing-destined and wingless-destined embryos can be collected.

Once both maternal tissue and embryonic tissue can be collected for the alternative morph inductions, how do we identify the mechanisms involved in the control

of genome expression? We know that both polyphenisms, wing and reproductive, require whole-body changes that include many transcriptional modulations. To get a clearer understanding of how polyphenic induction occurs, we propose the best place to start is at the genome level and take a top-down approach. This will include the generation of many large data sets such as transcriptomes, DNA methylomes, and chromatin structure and change mapping (open chromatin and specific histone marks). These types of data sets can then be integrated within a genome browser to facilitate further research once candidate regulators of the polyphenisms have been identified. In the absence of functional tools, a wide data integration approach will be valuable to narrow down the genomic regions involved in polyphenisms.

By using a top-down approach, we can recognize all likely regulatory candidates (regulatory regions) and investigate them further with more targeted genomic studies such as chromatin immunoprecipitation sequencing (ChIP-seq) to identify insulator, enhancer, or silencer regions or to assign function experimentally on a gene-by-gene basis (see Chaps. 6 and 7, this volume). Once candidate genes have been identified, tools like CRISPR and TALEs can be used for knockouts, knockdowns, and epigenetic modulation. Although there is still much to learn about phenotypic plasticity, the pea aphid is an excellent model for molecular studies of polyphenism that is just on the verge of an exponential increase of integrative genomic data.

References

1. West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford/New York
2. de Jong M, Leyser O (2012) Developmental plasticity in plants. Cold Spring Harb Symp Quant Biol 77:63–73. doi:10.1101/sqb.2012.77.014720
3. Aubin-Horth N, Renn SCP (2009) Genomic reaction norms: using integrative biology to understand molecular mechanisms of phenotypic plasticity. Mol Ecol 18(18):3763–3780. doi:10.1111/j.1365-294X.2009.04313.x
4. Beldade P, Mateus ARA, Keller RA (2011) Evolution and molecular mechanisms of adaptive developmental plasticity. Mol Ecol 20(7):1347–1363. doi:10.1111/j.1365-294X.2011.05016.x
5. De Loof A, Boerjan B, Ernst UR, Schoofs L (2013) The mode of action of juvenile hormone and ecdysone: towards an epi-endocrinological paradigm? Gen Comp Endocrinol 188:35–45. doi:10.1016/j.ygcen.2013.02.004
6. Bell O, Tiwari VK, Thoma NH, Schubeler D (2011) Determinants and dynamics of genome accessibility. Nat Rev Genet 12(8):554–564. doi:10.1038/nrg3017
7. Scheer BT (1961) The neuroendocrine system of arthropods. Vitam Horm 18:141–204. doi:10.1016/S0083-6729(08)60862-6
8. Tawfik AI (2012) Hormonal control of the phase polyphenism of the desert locust: a review of current understanding. Open Entomol J 6:22–41
9. Nijhout HF (1999) Control mechanisms of polyphenic development in insects: in polyphenic development, environmental factors alter some aspects of development in an orderly and predictable way. BioScience 49(3):181–192. doi:10.2307/1313508
10. Nijhout HF (1994) Insect hormones. Princeton University Press, Princeton

11. Kucharski R, Maleszka J, Foret S, Maleszka R (2008) Nutritional control of reproductive status in honeybees via DNA methylation. Science 319(5871):1827–1830. doi:10.1126/science.1153069
12. Foret S, Kucharski R, Pellegrini M, Feng S, Jacobsen SE, Robinson GE, Maleszka R (2012) DNA methylation dynamics, metabolic fluxes, gene splicing, and alternative phenotypes in honey bees. Proc Natl Acad Sci U S A 109(13):4968–4973. doi:10.1073/pnas.1202392109
13. Lyko F, Foret S, Kucharski R, Wolf S, Falckenhayn C, Maleszka R (2010) The honey bee epigenomes: differential methylation of brain DNA in queens and workers. PLoS Biol 8(11):e1000506. doi:10.1371/journal.pbio.1000506
14. Spannhoff A, Kim YK, Raynal NJ, Gharibyan V, Su MB, Zhou YY, Li J, Castellano S, Sbardella G, Issa JP, Bedford MT (2011) Histone deacetylase inhibitor activity in royal jelly might facilitate caste switching in bees. EMBO Rep 12(3):238–243. doi:10.1038/embor.2011.9
15. Simola DF, Ye C, Mutti NS, Dolezal K, Bonasio R, Liebig J, Reinberg D, Berger SL (2013) A chromatin link to caste identity in the carpenter ant *Camponotus floridanus*. Genome Res 23(3):486–496. doi:10.1101/gr.148361.112
16. Koch PB, Brakefield PM, Kesbeke F (1996) Ecdysteroids control eyespot size and wing color pattern in the polyphenic butterfly *Bicyclus anynana* (Lepidoptera: Satyridae). J Insect Physiol 42(3):223–230. doi:10.1016/0022-1910(95)00103-4
17. Wolschin F, Mutti NS, Amdam GV (2011) Insulin receptor substrate influences female caste development in honeybees. Biol Lett 7(1):112–115. doi:10.1098/rsbl.2010.0463
18. Snell-Rood EC, Moczek AP (2012) Insulin signaling as a mechanism underlying developmental plasticity: the role of *FOXO* in a nutritional polyphenism. PLoS One 7(4):e34857. doi:10.1371/journal.pone.0034857
19. Anstey ML, Rogers SM, Ott SR, Burrows M, Simpson SJ (2009) Serotonin mediates behavioral gregarization underlying swarm formation in desert locusts. Science 323(5914):627–630. doi:10.1126/science.1165939
20. Gilbert LI (2012) Insect endocrinology, 1st edn. Elsevier/Academic, London/Waltham
21. Blackman RL, Eastop VF (2000) Aphids on the world's crops: an identification and information guide, 2nd edn. Wiley, Chichester/New York
22. Dixon AFG (1974) Biology of aphids. J Entomol Ser A Gen Entomol 48(2):156–156. doi:10.1111/j.1365-3032.1974.tb00049.x
23. Oerke EC (1994) Crop production and crop protection: estimated losses in major food and cash crops. Elsevier, Amsterdam/New York
24. Morrison WP, Peairs F (1998) Response model concept and economic impact. Response model for an introduced pest—the Russian wheat aphid. Entomological Society of America, Lanham
25. Blackman R (1987) Reproduction, cytogenetics and development. Aphids: their biology, natural enemies, and control. Elsevier Science Publishers, Amsterdam
26. Smith MAH, MacKay PA (1989) Genetic variation in male alary dimorphism in populations of pea aphid, *Acyrthosiphon pisum*. Entomol Exp Appl 51(2):125–132. doi:10.1111/j.1570-7458.1989.tb01222.x
27. Braendle C, Caillaud MC, Stern DL (2005) Genetic mapping of aphicarus – a sex-linked locus controlling a wing polymorphism in the pea aphid (*Acyrthosiphon pisum*). Heredity 94(4):435–442. doi:10.1038/sj.hdy.6800633
28. Müller CB, Williams IS, Hardie J (2001) The role of nutrition, crowding and interspecific interactions in the development of winged aphids. Ecol Entom 26(3):330–340. doi:10.1046/j.1365-2311.2001.00321.x
29. Zera AJ, Denno RF (1997) Physiology and ecology of dispersal polymorphism in insects. Annu Rev Entomol 42:207–230. doi:10.1146/annurev.ento.42.1.207
30. Braendle C, Davis GK, Brisson JA, Stern DL (2006) Wing dimorphism in aphids. Heredity 97(3):192–199. doi:10.1038/sj.hdy.6800863
31. Mueller WC, Rochow WF (1961) An aphid-injection method for the transmission of barley yellow dwarf virus. Virology 14:253–258

32. Kring JB (1977) Structure of the eyes of the pea aphid, *Acyrthosiphon pisum*. Ann Entomol Soc Am 70(6):855–860
33. Bromley AK, Dunn JA, Anderson M (1979) Ultrastructure of the antennal sensilla of aphids. I. Coeloconic and placoid sensilla. Cell Tissue Res 203(3):427–442
34. Lyko F, Maleszka R (2011) Insects as innovative models for functional studies of DNA methylation. Trends Genet 27(4):127–131. doi:10.1016/j.tig.2011.01.003
35. Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481–514. doi:10.1146/annurev.biochem.74.010904.153721
36. Probst AV, Dunleavy E, Almouzni G (2009) Epigenetic inheritance during the cell cycle. Nat Rev Mol Cell Biol 10(3):192–206. doi:10.1038/nrm2640
37. Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9(6):465–476. doi:10.1038/nrg2341
38. Jones PA (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. Nat Rev Genet 13(7):484–492. doi:10.1038/nrg3230
39. Fournier A, Sasai N, Nakao M, Defossez P-A (2012) The role of methyl-binding proteins in chromatin organization and epigenome maintenance. Brief Funct Genomics 11(3):251–264. doi:10.1093/bfgp/elr040
40. Walsh TK, Brisson JA, Robertson HM, Gordon K, Jaubert-Possamai S, Tagu D, Edwards OR (2010) A functional DNA methylation system in the pea aphid, *Acyrthosiphon pisum*. Insect Mol Biol 19(Suppl 2):215–228. doi:10.1111/j.1365-2583.2009.00974.x
41. Hellman A, Chess A (2007) Gene body-specific methylation on the active X chromosome. Science 315(5815):1141–1143. doi:10.1126/science.1136352
42. Shukla S, Kavak E, Gregory M, Imashimizu M, Shutinoski B, Kashlev M, Oberdoerffer P, Sandberg R, Oberdoerffer S (2011) CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. Nature 479(7371):74–79. doi:10.1038/nature10442
43. Hunt BG, Glastad KM, Yi SV, Goodisman MA (2013) The function of intragenic DNA methylation: insights from insect epigenomes. Integr Comp Biol 53(2):319–328. doi:10.1093/icb/ict003
44. Bonasio R, Li Q, Lian J, Mutti Navdeep S, Jin L, Zhao H, Zhang P, Wen P, Xiang H, Ding Y, Jin Z, Shen Steven S, Wang Z, Wang W, Wang J, Berger Shelley L, Liebig J, Zhang G, Reinberg D (2012) Genome-wide and caste-specific DNA methylomes of the ants *Camponotus floridanus* and *Harpegnathos saltator*. Curr Biol 22(19):1755–1764. doi:10.1016/j.cub.2012.07.042
45. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee BK, Sheffield NC, Graf S, Huss M, Keefe D, Liu Z, London D, McDaniell RM, Shibata Y, Showers KA, Simon JM, Vales T, Wang T, Winter D, Zhang Z, Clarke ND, Birney E, Iyer VR, Crawford GE, Lieb JD, Furey TS (2011) Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res 21(10):1757–1767. doi:10.1101/gr.121541.111
46. Van Holde KE (1989) Chromatin. Springer series in molecular biology. Springer, New York
47. Ho L, Crabtree GR (2010) Chromatin remodelling during development. Nature 463(7280):474–484. doi:10.1038/nature08911
48. Zhou VW, Goren A, Bernstein BE (2011) Charting histone modifications and the functional organization of mammalian genomes. Nat Rev Genet 12(1):7–18. doi:10.1038/nrg2905
49. Rider SD Jr, Srinivasan DG, Hilgarth RS (2010) Chromatin-remodelling proteins of the pea aphid, *Acyrthosiphon pisum* (Harris). Insect Mol Biol 19(Suppl 2):201–214. doi:10.1111/j.1365-2583.2009.00972.x
50. Mandrioli M, Azzoni P, Lombardo G, Manicardi GC (2011) Composition and epigenetic markers of heterochromatin in the aphid *Aphis nerii* (Hemiptera: Aphididae). Cytogenet Genome Res 133(1):67–77
51. Mandrioli M, Borsatti F (2007) Analysis of heterochromatic epigenetic markers in the holocentric chromosomes of the aphid *Acyrthosiphon pisum*. Chromosome Res 15(8):1015–1022. doi:10.1007/s10577-007-1176-4
52. Hawkins RD, Hon GC, Ren B (2010) Next-generation genomics: an integrative approach. Nat Rev Genet 11(7):476–486. doi:10.1038/nrg2795

53. Fraga MF, Esteller M (2002) DNA methylation: a profile of methods and applications. BioTechniques 33(3):632, 634, 636–649
54. McKay DJ, Lieb JD (2013) A common set of DNA regulatory elements shapes Drosophila appendages. Dev Cell 27(3):306–318. doi:10.1016/j.devcel.2013.10.009
55. The International Aphid Genomics C (2010) Genome sequence of the pea aphid (*Acyrthosiphon pisum*). PLoS Biol 8(2):e1000313. doi:10.1371/journal.pbio.1000313
56. Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispe C, Collin O, Richards S, Wilson ACC, Murphy T, Tagu D (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol 19:5–12. doi:10.1111/j.1365-2583.2009.00930.x
57. Zhou L, Cheng X, Connolly BA, Dickman MJ, Hurd PJ, Hornby DP (2002) Zebularine: a novel DNA methylation inhibitor that forms a covalent complex with DNA methyltransferases. J Mol Biol 321(4):591–599
58. Zhou P, Lu Y, Sun XH (2011) Zebularine suppresses TGF-beta-induced lens epithelial cell-myofibroblast transdifferentiation by inhibiting MeCP2. Mol Vis 17:2717–2723
59. Cheng JC, Matsen CB, Gonzales FA, Ye W, Greer S, Marquez VE, Jones PA, Selker EU (2003) Inhibition of DNA methylation and reactivation of silenced genes by zebularine. J Natl Cancer Inst 95(5):399–409
60. Dombrovsky A, Arthaud L, Ledger TN, Tares S, Robichon A (2009) Profiling the repertoire of phenotypes influenced by environmental cues that occur during asexual reproduction. Genome Res 19(11):2052–2063. doi:10.1101/gr.091611.109
61. Mutti NS, Park Y, Reese JC, Reeck GR (2006) RNAi knockdown of a salivary transcript leading to lethality in the pea aphid, *Acyrthosiphon pisum*. J Insect Sci 6(38):1–7. doi:10.1673/031.006.3801
62. Jaubert-Possamai S, Le Trionnaire G, Bonhomme J, Christophides GK, Rispe C, Tagu D (2007) Gene knockdown by RNAi in the pea aphid *Acyrthosiphon pisum*. BMC Biotechnol 7:63. doi:10.1186/1472-6750-7-63
63. Christiaens O, Swevers L, Smagghe G (2014) DsRNA degradation in the pea aphid (*Acyrthosiphon pisum*) associated with lack of response in RNAi feeding and injection assay. Peptides 53:307–314. doi:10.1016/j.peptides.2013.12.014
64. Bhaya D, Davison M, Barrangou R (2011) CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. Annu Rev Genet 45:273–297. doi:10.1146/annurev-genet-110410-132430
65. Terns MP, Terns RM (2011) CRISPR-based adaptive immune systems. Curr Opin Microbiol 14(3):321–327. doi:10.1016/j.mib.2011.03.005
66. Wiedenheft B, Sternberg SH, Doudna JA (2012) RNA-guided genetic silencing systems in bacteria and archaea. Nature 482(7385):331–338. doi:10.1038/nature10886
67. Pennisi E (2013) The CRISPR craze. Science 341(6148):833–836. doi:10.1126/science.341.6148.833
68. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337(6096):816–821. doi:10.1126/science.1225829
69. Terns RM, Terns MP (2014) CRISPR-based technologies: prokaryotic defense weapons repurposed. Trends Genet 30(3):111–118. doi:10.1016/j.tig.2014.01.003
70. Hsu Patrick D, Lander Eric S, Zhang F (2014) Development and applications of CRISPR-Cas9 for genome engineering. Cell 157(6):1262–1278. doi:10.1016/j.cell.2014.05.010
71. Rusk N (2014) CRISPRs and epigenome editing. Nat Methods 11(1):28
72. Friedland AE, Tzur YB, Esvelt KM, Colaiacovo MP, Church GM, Calarco JA (2013) Heritable genome editing in *C. elegans* via a CRISPR-Cas9 system. Nat Methods 10(8):741–743. doi:10.1038/nmeth.2532
73. Hwang WY, Fu Y, Reyon D, Maeder ML, Tsai SQ, Sander JD, Peterson RT, Yeh JR, Joung JK (2013) Efficient genome editing in zebrafish using a CRISPR-Cas system. Nat Biotechnol 31(3):227–229. doi:10.1038/nbt.2501

74. Joung JK, Sander JD (2013) TALENs: a widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14(1):49–55. doi:10.1038/nrm3486
75. Maeder ML, Angstman JF, Richardson ME, Linder SJ, Cascio VM, Tsai SQ, Ho QH, Sander JD, Reyon D, Bernstein BE, Costello JF, Wilkinson MF, Joung JK (2013) Targeted DNA demethylation and activation of endogenous genes using programmable TALE-TET1 fusion proteins. Nat Biotechnol 31(12):1137–1142. doi:10.1038/nbt.2726
76. Mendenhall EM, Williamson KE, Reyon D, Zou JY, Ram O, Joung JK, Bernstein BE (2013) Locus-specific editing of histone modifications at endogenous enhancers. Nat Biotechnol 31(12):1133–1136. doi:10.1038/nbt.2701
77. Shi Y, Lan F, Matson C, Mulligan P, Whetstine JR, Cole PA, Casero RA, Shi Y (2004) Histone demethylation mediated by the nuclear amine oxidase homolog LSD1. Cell 119(7):941–953. doi:10.1016/j.cell.2004.12.012
78. Liang G, Lin JCY, Wei V, Yoo C, Cheng JC, Nguyen CT, Weisenberger DJ, Egger G, Takai D, Gonzales FA, Jones PA (2004) Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. Proc Natl Acad Sci U S A 101(19):7357–7362. doi:10.1073/pnas.0401866101
79. Zhou P, Lu Y, Sun XH (2012) Effects of a novel DNA methyltransferase inhibitor Zebularine on human lens epithelial cells. Mol Vis 18:22–28
80. Larson MH, Gilbert LA, Wang X, Lim WA, Weissman JS, Qi LS (2013) CRISPR interference (CRISPRi) for sequence-specific control of gene expression. Nat Protoc 8(11):2180–2196. doi:10.1038/nprot.2013.132
81. Ghanim M, Dombrovsky A, Raccah B, Sherman A (2006) A microarray approach identifies ANT, OS-D and takeout-like genes as differentially regulated in alate and apterous morphs of the green peach aphid *Myzus persicae* (Sulzer). Insect Biochem Mol Biol 36(11):857–868. doi:10.1016/j.ibmb.2006.08.007
82. Brisson JA, Davis GK, Stern DL (2007) Common genome-wide patterns of transcript accumulation underlying the wing polyphenism and polymorphism in the pea aphid (*Acyrthosiphon pisum*). Evol Dev 9(4):338–346. doi:10.1111/j.1525-142X.2007.00170.x
83. Purandare SR, Bickel RD, Jaquiery J, Rispe C, Brisson JA (2014) Accelerated evolution of morph-biased genes in pea aphids. Mol Biol Evol 31(8):2073–2083. doi:10.1093/molbev/msu149
84. Liu L-J, Zheng H-Y, Jiang F, Guo W, Zhou S-T (2014) Comparative transcriptional analysis of asexual and sexual morphs reveals possible mechanisms in reproductive polyphenism of the cotton aphid. PLoS One 9(6):e99506. doi:10.1371/journal.pone.0099506
85. Cortes T, Tagu D, Simon JC, Moya A, Martinez-Torres D (2008) Sex versus parthenogenesis: a transcriptomic approach of photoperiod response in the model aphid *Acyrthosiphon pisum* (Hemiptera: Aphididae). Gene 408(1–2):146–156. doi:10.1016/j.gene.2007.10.030
86. Le Trionnaire G, Francis F, Jaubert-Possamai S, Bonhomme J, De Pauw E, Gauthier JP, Haubruge E, Legeai F, Prunier-Leterme N, Simon JC, Tanguy S, Tagu D (2009) Transcriptomic and proteomic analyses of seasonal photoperiodism in the pea aphid. BMC Genomics 10:456. doi:10.1186/1471-2164-10-456
87. Steel CG, Lees AD (1977) The role of neurosecretion in the photoperiodic control of polymorphism in the aphid Megoura viciae. J Exp Biol 67:117–135
88. Le Trionnaire G, Jaubert-Possamai S, Bonhomme J, Gauthier JP, Guernec G, Le Cam A, Legeai F, Monfort J, Tagu D (2012) Transcriptomic profiling of the reproductive mode switch in the pea aphid in response to natural autumnal photoperiod. J Insect Physiol 58(12):1517–1524. doi:10.1016/j.jinsphys.2012.07.009
89. Le Trionnaire G, Jaubert S, Sabater-Munoz B, Benedetto A, Bonhomme J, Prunier-Leterme N, Martinez-Torres D, Simon JC, Tagu D (2007) Seasonal photoperiodism regulates the expression of cuticular and signalling protein genes in the pea aphid. Insect Biochem Mol Biol 37(10):1094–1102. doi:10.1016/j.ibmb.2007.06.008

90. Gallot A, Shigenobu S, Hashiyama T, Jaubert-Possamai S, Tagu D (2012) Sexual and asexual oogenesis require the expression of unique and shared sets of genes in the insect *Acyrthosiphon pisum*. BMC Genomics 13:76. doi:10.1186/1471-2164-13-76
91. Le Trionnaire G, Wucher V, Tagu D (2013) Genome expression control during the photoperiodic response of aphids. Physiol Entomol 38(2):117–125. doi:10.1111/phen.12021
92. Gallot A, Rispe C, Leterme N, Gauthier J-P, Jaubert-Possamai S, Tagu D (2010) Cuticular proteins and seasonal photoperiodism in aphids. Insect Biochem Mol Biol 40(3):235–240. doi:10.1016/j.ibmb.2009.12.001
93. Ma Z, Guo W, Guo X, Wang X, Kang L (2011) Modulation of behavioral phase changes of the migratory locust by the catecholamine metabolic pathway. Proc Natl Acad Sci 108(10):3882–3887. doi:10.1073/pnas.1015098108
94. Hardie JIM, Lees AD (1985) The induction of normal and teratoid viviparae by a juvenile hormone and kinoprene in two species of aphids. Physiol Entomol 10(1):65–74. doi:10.1111/j.1365-3032.1985.tb00020.x
95. Hardie J, Gao N, Timar T, Sebok P, Honda K (1996) Precocene derivatives and aphid morphogenesis. Arch Insect Biochem Physiol 32(3–4):493–501. doi:10.1002/(sici)1520-6327(1996)32:3/4<493::aid-arch21>3.0.co;2-6
96. Ishikawa A, Ishikawa Y, Okada Y, Miyazaki S, Miyakawa H, Koshikawa S, Brisson JA, Miura T (2012) Screening of upregulated genes induced by high density in the vetch aphid *Megoura crassicauda*. J Exp Zool A Ecol Genet Physiol 317(3):194–203. doi:10.1002/jez.1713
97. Schwedes CC, Carney GE (2012) Ecdysone signaling in adult *Drosophila melanogaster*. J Insect Physiol 58(3):293–302. doi:10.1016/j.jinsphys.2012.01.013
98. Miura T, Braendle C, Shingleton A, Sisk G, Kambhampati S, Stern DL (2003) A comparison of parthenogenetic and sexual embryogenesis of the pea aphid *Acyrthosiphon pisum* (Hemiptera: Aphidoidea). J Exp Zool B Mol Dev Evol 295(1):59–81. doi:10.1002/jez.b.3
99. Ishikawa A, Miura T (2013) Transduction of high-density signals across generations in aphid wing polyphenism. Physiol Entomol 38(2):150–156. doi:10.1111/phen.12022

Chapter 6
Insect Regulatory Genomics

Kushal Suryamohan and Marc S. Halfon

Abstract Insects are the most diverse and ecologically important group of animals in the animal kingdom, with more than a million species described to date. Whole-genome sequencing, which has revolutionized many areas of biological research, carries significant potential for achieving a deeper understanding of insect development, physiology, and evolution and for facilitating new biotechnological advances in insect management and biocontrol. Comprehensive genome annotation, including not only genes but also regulatory regions, is necessary for realizing the full benefits of this sequencing. However, regulatory element discovery in non-model organisms remains a major challenge as most regulatory sequences have diverged past the point of recognition by standard sequence alignment methods, even for relatively closely related species such as flies and mosquitoes. We review here some of the advances made in insect regulatory genomics and the methods and resources available for identifying regulatory elements in well-studied model insects such as *Drosophila*. We discuss recent efforts to extend these approaches to discovering

K. Suryamohan
Department of Biochemistry, University at Buffalo-State University of New York,
701 Ellicott St, Buffalo, NY 14203, USA

NY State Center of Excellence in Bioinformatics and Life Sciences,
Buffalo, NY 14203, USA
e-mail: kushalsuryamohan@gmail.com

M.S. Halfon (✉)
Department of Biochemistry, University at Buffalo-State University of New York,
701 Ellicott St, Buffalo, NY 14203, USA

Department of Biological Sciences, University at Buffalo-State University of New York,
701 Ellicott St, Buffalo, NY 14203, USA

Department of Biomedical Informatics, University at Buffalo-State University of New York,
701 Ellicott St, Buffalo, NY 14203, USA

Program in Genetics, Genomics and Bioinformatics, University at Buffalo-State University of New York, 701 Ellicott St, Buffalo, NY 14203-1101, USA

NY State Center of Excellence in Bioinformatics and Life Sciences,
Buffalo, NY 14203, USA

Molecular and Cellular Biology Department and Program in Cancer Genetics,
Roswell Park Cancer Institute, Buffalo, NY 14263, USA
e-mail: mshalfon@buffalo.edu

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*,
Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_6

regulatory elements in evolutionarily diverged non-model species and potential applications of the resulting regulatory data.

Abbreviations

B1H	Bacterial one-hybrid
Cas9	CRISPR-associated protein 9
ChIP	Chromatin immunoprecipitation
ChIP–chip	Chromatin immunoprecipitation combined with genome-tiling microarrays
ChIP-seq	Chromatin immunoprecipitation combined with next-generation sequencing
CRISPR	Clustered regularly interspaced short palindromic repeats
CRM	*cis*-regulatory module
DNase-seq	DNase I digestion combined with sequencing
DPE	Downstream promoter element
dsRNA	Double-stranded RNA
FACS	Fluorescently activated cell sorting
FAIRE-seq	Formaldehyde-assisted isolation of regulatory elements
GFP	Green fluorescent protein
GMOD	Generic Model Organism Database
GTF	General transcription factor
MOD	Model organism database
NCBI	National Center for Biotechnology Information
PBM	Protein-binding microarray
PWM	Position weight matrix
RNA-seq	RNA sequencing
RNAi	RNA interference
STARR-seq	Self-transcribing active regulatory region sequencing
TALENS	Transcription activator-like effector nucleases
TF	Transcription factor
TFBS	Transcription factor binding site
ZFN	Zinc finger nuclease

6.1 The Importance of Regulatory DNA: Why Regulate Genes?

The expression of metazoan protein-coding genes is regulated at several steps in the pathway from DNA to protein, including transcription of DNA to mRNA; mRNA stability, transport, processing, and translation; and posttranslational protein

modification. Such stratified control allows cells exquisite control over which proteins they make, and this confers distinct properties to cells, resulting in cell differentiation and diversity.

The main mechanism by which control of gene expression is achieved is transcriptional regulation. Although a promoter is necessary to initiate gene transcription, a significant part of eukaryotic transcriptional regulation is mediated by distal *cis*-regulatory modules (CRMs), of which the most common forms are known as enhancers: clusters of transcription factor binding sites (TFBS) that act without regard to orientation, distance, or location (up- or downstream) relative to the transcribed gene [1]. Regulation of gene expression is also achieved by additional distal *cis*-acting regulatory elements that include silencers, insulators, and locus control regions.

Next-generation DNA sequencing technologies now enable us to sequence the genomes of many organisms in their entirety relatively rapidly and at constantly decreasing cost. These technological developments have made possible numerous insect genome projects, many of which have now been completed and many more of which are anticipated: the i5K project aims to sequence 5000 insect and other arthropod genomes over the next 5 years [2]. The sequence of a genome, however, is of limited use without its annotation. That is, in addition to the DNA sequence, it is necessary to attach biological information to a genome, including not only identifying protein-coding genes and their coding exons but also defining non-protein-coding genes and, crucially, the different aforementioned regulatory elements—and then assigning function to each.

Historically, annotation of regulatory regions has been a challenge even in well-studied model organisms due to the inherent difficulties involved in regulatory sequence identification. In non-model organisms, where there are few experimental genetic and molecular data, where little is known about most transcription factors and their target binding sites, and where the ability to make transgenic animals is severely limited, genome annotation is particularly difficult. However, the extensive advances achieved by virtue of *Drosophila*'s position as a leading model organism have laid the foundation for molecular and computational tools to study other arthropods. Thus, annotating regulatory regions in other insect species is now becoming a realistic task, one that is essential to understanding the development and physiology of insects and which carries the potential to enable the development of products and techniques to control the harmful aspects of insects on society, as well as to harness the many benefits we derive from them. In this chapter, we discuss salient studies in insect regulatory genomics, focusing mainly on methods developed to identify regulatory elements, and highlight a few key studies on regulatory elements in insects.

6.2 Regulatory Genomic Analysis in Insects

6.2.1 Gene Function

Currently, there are about 200 insect species whose genomes have been sequenced or have begun to be sequenced. Details of each sequencing project and genome data are available at the National Center for Biotechnology Information (NCBI) Entrez Genome Project webpage (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi) and the i5k website (http://www.arthropodgenomes.org/wiki/i5K). Finished insect genome projects include *Drosophila melanogaster* as well as 19 other *Drosophila* species (http://www.flybase.org, ftp://ftp.hgsc.bcm.edu/Dmelanogaster/); medically important species that are vectors for diseases such as malaria (*Anopheles gambiae*), dengue fever (*Aedes aegypti*), and elephantiasis (*Culex pipiens*); agriculturally important species such as the honeybee (*Apis mellifera*) and silkworm (*Bombyx mori*); pests such as the red flour beetle (*Tribolium castaneum*) and the pea aphid (*Acyrthosiphon pisum*) (see Chaps. 4 and 5, in this volume); the parasitoid jewel wasp (*Nasonia vitripennis*); several species of ants and butterflies (see Chap. 3, in this volume); and others. The availability of the sequenced genomes of these insects, combined with the efforts of a diverse group of researchers, has dramatically improved the molecular and genetic tools available to study them with the consequence that many of these are now considered model or emerging-model research organisms.

While a significant proportion of the genes so far identified in insect genomes have been assigned a known or putative function (largely through homology to known genes in *Drosophila* and other organisms), many genes have yet to reveal their function. Fortunately, recent years have seen a surge in methods for the study of formerly non-model insect species, including gene knockdowns by RNA interference and genome engineering using transcription activator-like effector nucleases (TALENs) [3] or the clustered regularly interspaced short palindromic repeats/CRISPR-associated protein 9 (CRISPR/Cas9) system [4]. The discovery of transposons such as *piggyBac* [5] and *Hermes* [6] in the moth *Trichoplusia ni* and the house fly *Musca domestica*, respectively, has now made it possible to perform gene perturbation studies in a wide range of insects through the development of transgenic technology (see Sect. 6.3) [7–12].

As with many technologies, the use of RNA interference (RNAi) in insects was pioneered in *Drosophila* [13], but this powerful method was rapidly applied to the red flour beetle *T. castaneum* and many other holometabolous insect species [14, 15], including the malaria vector mosquito *Anopheles gambiae* [16, 17]. The development of parental RNAi techniques [18] meant that gene function in embryos could be disrupted by injecting double-stranded RNA (dsRNA) into pupal or adult females [19], allowing for studies of early insect development [20]. The immense utility of RNAi was also realized when it was used to study *Hox* gene function for the first time in a hemimetabolous species, the bug *Oncopeltus fasciatus* [21]. For an overview of successful applications of RNAi technology to assign functions to genes in various insects, readers are referred to the excellent review article by Xavier Belles [22].

With the recent development of methods such as zinc finger nucleases (ZFNs), TALENs, and CRISPR/Cas9 [3, 23–25], genome editing has become feasible in a broad range of species [26, 27], resulting in a deluge of papers despite the first reported uses of CRISPR taking place only a year ago. While as usual the first applications of CRISPR/Cas9 technology in insects have been in *Drosophila* [24, 25, 28], its success there has now encouraged its use in other insect species [29], and this powerful system is likely to revolutionize experimental studies in model and non-model insects alike. With these new technologies in hand, it is only a matter of time before we have a better understanding of the functions of genes crucial to development, vector biology, and other biological processes in most insect species.

6.2.2 Discovering DNA Regulatory Elements in the Genome

The task of finding regulatory elements in the genome has historically been a challenging one. Approaches can be classified into two broad categories: empirical and computational. Empirical approaches (Fig. 6.1) have traditionally been time-consuming, labor-intensive, and expensive. The genomic era has brought about the development of genome-wide, high-throughput assays and techniques, which has greatly accelerated the pace of regulatory element discovery. However, these methods are not without limitations: they can remain very costly, are often difficult to validate, and typically do not produce comprehensive or fully accurate results. This is a particular problem for *cis*-regulatory modules (CRMs), which may be functional only in certain cell types or under specific conditions.

Computational methods (Fig. 6.2) have provided an attractive complementary approach for regulatory element identification. However, these methods too have drawbacks, including high false-positive prediction rates and the challenges of large-scale empirical validation. Despite this, significant advances have been made in the computational methods for modeling and detection of DNA regulatory elements over the last decade. The availability of complete genome sequences for multiple organisms, whole-transcriptome profiles, high-throughput experimental methods for mapping protein-binding sites in DNA, increased throughput in empirical identification of CRMs, elucidation of higher order structures of the regulatory sequences [30, 31], and more efficient assays for testing putative regulatory regions have all contributed to the development of successful methods. Nevertheless, these approaches have primarily been limited to a few well-understood model organisms and biological systems, where a fair amount of prior knowledge is available, where the organisms are amenable to experimental manipulation, and where there is a large community-driven funding base.

In this section, we will briefly summarize the different empirical and computational approaches used for regulatory element discovery with particular focus on the identification of regulatory elements in insect species (including *D. melanogaster*).

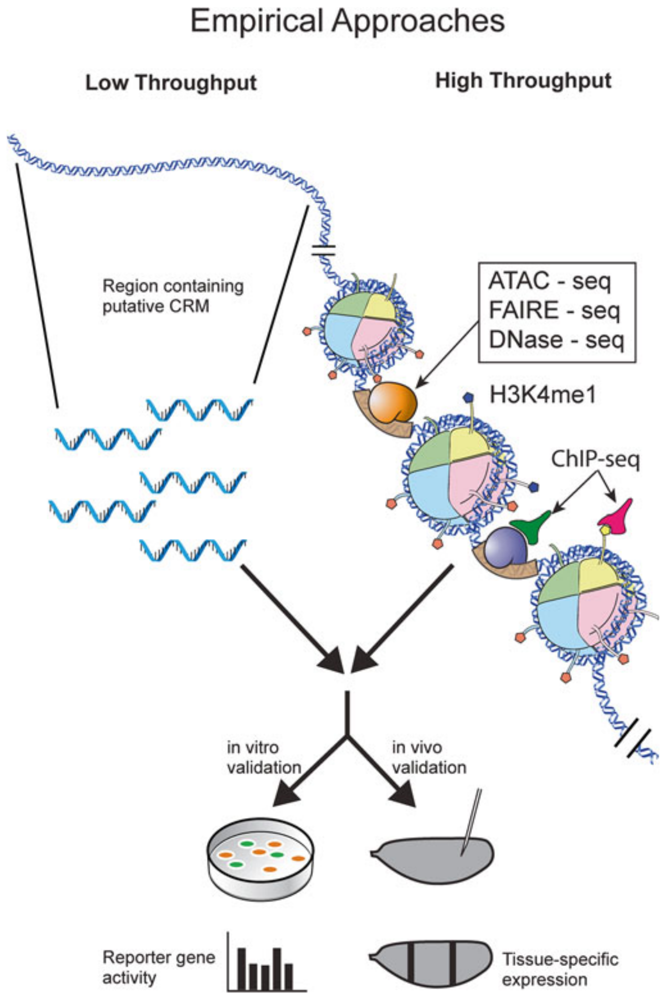


Fig. 6.1 Empirical approaches to CRM discovery. Empirical approaches to CRM discovery can be broadly classified into *low-throughput* (*left*) and *high-throughput* (*right*) methods. In both cases, the putative CRMs are tested in a heterologous reporter system. Low-throughput methods involve testing of isolated regions of DNA that contain putative CRMs in a cell culture or transgenic animal setting (or both). In the former, putative CRMs are transfected into cultured cells and reporter gene activity (e.g., luciferase, GFP) levels are quantified relative to a control vector. In the latter experiment, transgenic animals (here, flies) bearing a reporter gene construct are generated and assayed for tissue-specific expression patterns driven by the putative CRM. High-throughput methods make use of next-generation sequencing to identify potential regions of regulatory DNA. Chromatin immunoprecipitation-based methods use antibodies to detect binding of TFs of interest genome-wide followed by sequencing to identify the bound regions (ChIP-seq). A variant of this uses antibodies against specific chromatin modifications, such as histone methylation (e.g., H3K4me1), that are characteristic of regulatory sequences. A third variant makes use of the fact that regulatory regions have an "open" chromatin configuration, i.e., are depleted of nucleosomes or otherwise more accessible to cleavage by DNase I (DNase-seq) or to transposon insertion (ATAC-seq), or respond differently to chemical fractionation (FAIRE-seq). Additional methods are discussed in the text. Like predictions from low-throughput approaches, results from high-throughput experiments can also be tested using cell culture or transgenic methods, although typically only a fraction of the predictions can be validated

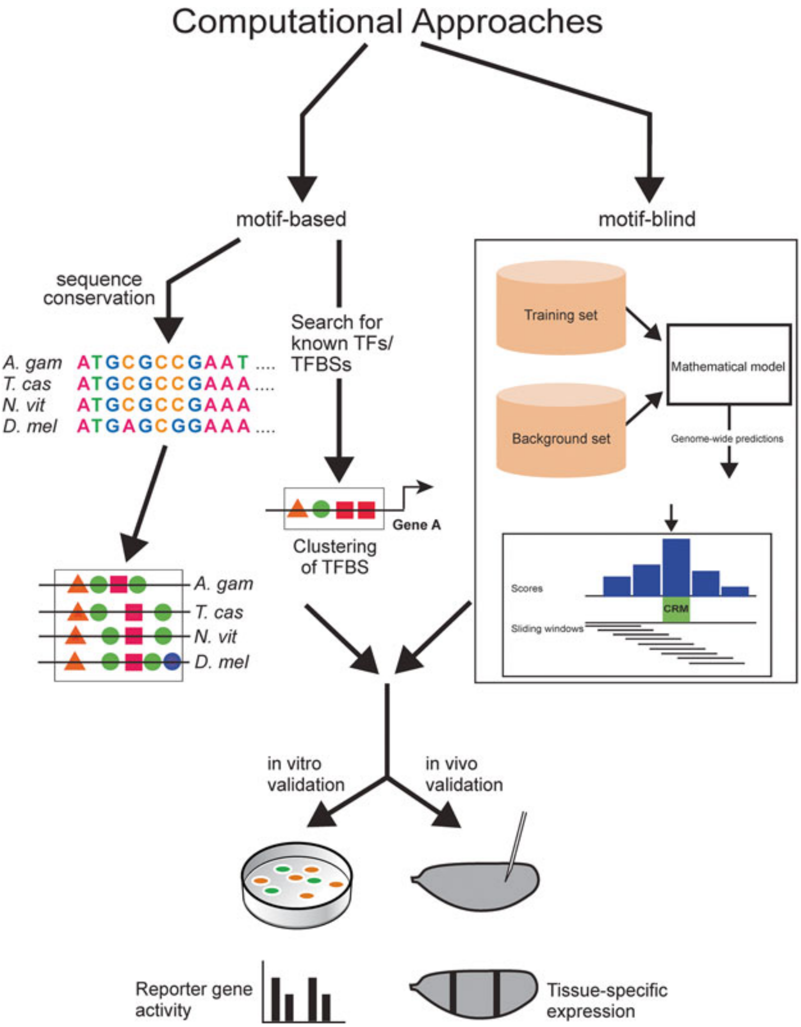


Fig. 6.2 Computational approaches to CRM discovery. Computational approaches can be broadly classified into *motif-based* (*left*) and *motif-blind* (*right*) methods. In the former, all observed instances of a TFBS (which are usually short motifs of ~6–10 bp) are modeled into a position weight matrix (PWM – see Sect. 6.2.2.2). Motif-based methods are predicated upon the knowledge that CRMs consist of clusters of TFBSs in a small region of DNA; these clusters can be searched genome wide (*center*). *Sequence conservation-based* methods (*left*) look for evolutionarily constrained regions of noncoding DNA containing clusters of TFBSs across several closely or distantly related species. The colored triangles, squares, and circles each represent a specific instance of a particular TFBS in a segment of DNA. Motif-blind methods (*right*) are unique in that they do not rely on existing knowledge of TFBSs or TFs and instead make use of the statistical profiles of experimentally validated CRMs (the "training set") against a set of non-CRMs (the "background set"). A statistical model is then used to scan the whole genome of a candidate species using overlapping windows with a score assigned to each window; the highest peaks in the resulting score profile are predicted to be CRMs. As with empirical approaches, predictions from computational methods can then be tested using a variety of cell culture or transgenic validation methods

6.2.2.1 Empirical Discovery of Regulatory Regions

Promoters

Initiation of transcription is achieved by the promoter, which can be viewed as consisting of a core promoter along with a variable number of proximal promoter elements (for variants of promoter architecture, refer to [32]). Together, these regions integrate regulatory inputs and initiate gene transcription. The core promoter consists of TFBSs for general transcription factors (GTFs) necessary to recruit RNA polymerase II (reviewed in [33]) and is typically defined as the ~40 bp region on either side of the transcriptional start site of its gene. While a number of core promoter binding motifs have been defined (e.g., the familiar TATA box and the downstream promoter element (DPE) [34]), there are no universal motifs common to all promoters, and the majority of promoters do not appear to contain any of the well-characterized motifs [35].

In the last decade, several high-throughput next-generation sequencing-based methods have been developed to aid promoter identification, including capture and sequencing of the 5′ ends of mRNA transcripts (CAGE-seq [36], PEAT [37], RAMPAGE [38]) and chromatin immunoprecipitation (ChIP)-based methods (e.g., ChIP of RNA Pol II) ([39–42]; see also Chap. 7, in this volume). In insects, genome-wide characterization of promoters has largely been restricted to *D. melanogaster*, an issue that needs to be addressed for other emerging-model or non-model insect species whose genomes have been sequenced [37, 38, 43, 44].

Cis-Regulatory Modules (CRMs)

Traditional CRM Discovery Methods

Whereas promoters can be identified through capture of 5′ mRNA sequences or RNA Pol II binding, and to a lesser extent by virtue of the presence of defined sequence motifs, discovery of distal CRMs presents a much greater challenge. Unlike promoters, CRMs do not contain broadly recognizable sequence characteristics and do not lend themselves to discovery via simple transcriptional profiling-based methods. Empirical approaches to discovering enhancers have historically involved isolating fragments of DNA containing putative CRMs and cloning them upstream of a minimal promoter fused to a reporter gene to test for transcriptional activity in cell lines or transgenic animals. Although more laborious and expensive to conduct than cell culture assays, transgenic animal studies have the great advantage of providing spatiotemporal expression information simultaneously in all tissues and cell types of an overall wild-type animal. Early empirical approaches were limited in the number of assays that could feasibly be performed. However, the more recent sequencing of the genomes of multiple species, along with the availability of next-generation sequencing strategies, has allowed for the development of higher-throughput methods for regulatory element identification in model organisms such as *Drosophila* and mouse, resulting in an explosion of newly predicted—and in many cases validated—CRMs.

Two efforts in *Drosophila* are notable for both their scope and audacity. Groups at the Howard Hughes Medical Institute's Janelia Farm Research Campus and at the Research Institute of Molecular Pathology in Vienna have taken a genome-tiling approach in which short overlapping segments of noncoding DNA are assayed in a more-or-less unbiased fashion using in vivo reporter gene assays. Collectively, these two groups have generated some 14,000 new reporter lines, increasing in the last few years by 5–7-fold the cumulative efforts of the preceding three decades [45, 46]. It should be noted, however, that many of the tested sequences are on the order of 2–3 kb and as such may contain multiple CRMs (which are frequently less than 500 bp in length). Thus, precise mapping of individual regulatory elements may still require substantial follow-up.

TFBS Discovery

Although such massive undertakings seem an unlikely prospect for extension to other insect species, the rise of microarrays and next-generation sequencing has spawned a growing number of high-throughput yet more broadly accessible methods for both TFBS and CRM discovery. Sensitive, unbiased methods to identify and characterize TFBSs in a systematic manner include SELEX-seq [47], protein-binding microarrays (PBMs) [48], and large-scale bacterial one-hybrid (B1H) assays. The latter are especially advantageous as they can determine the specificities of a TF of interest without requiring purification of the TF [49]. Chromatin immunoprecipitation (ChIP) coupled with genome-tiling microarrays (ChIP–chip [42]), now largely supplanted by next-generation sequencing (ChIP-seq) [39], enables genome-wide identification of regions bound in vivo by a given transcription factor (TF). Regions isolated from ChIP-based assays usually range in size from a few dozen base pairs to a few hundred base pairs. Since the regions obtained from ChIP are larger than the actual TFBSs themselves, additional computational analysis is needed to discover the individual TFBS within these regions. These limitations can be overcome with application of newer methods such as ChIP-exo (in which an exonuclease trims the DNA to give a higher resolution in TFBS mapping) [50] or extremely deep sequencing, which can reveal transcription factor binding sites 10–20 bp long ("digital footprinting") [51]. As always, it is worth bearing in mind the caveat that it cannot always be certain that all observed protein–DNA interactions have an active role in regulation. In at least some instances, substantial in vivo binding has been detected at sequences that do not appear to have regulatory function, and the number of sites bound by a TF can greatly exceed the number of genes the TF is believed to regulate [52–54]. Binding is also cell type specific, meaning that ChIP-based methods are most effective when applied to pure cell populations and provide more limited information when performed on complex tissues or whole embryos. Nevertheless, sufficient data to make reasonable inferences as to probable binding of a given TF at a given locus, through collective application of the discussed approaches, are likely within reach for the majority of TFs in *Drosophila* in the near future. Since TF binding domains have frequently evolved slowly overall [55], in many cases extrapolation to other insect species will also be possible.

CRM Discovery Using Epigenomic Methods

Active regulatory regions tend to be devoid of nucleosomes, a property that can be exploited for regulatory element discovery. Regions of nucleosome-depleted, or "open," chromatin can be identified on a genome-wide scale through methods such as DNase-seq [56], where accessible regions are detected by virtue of higher susceptibility to enzymatic cleavage by DNase I; FAIRE-seq (formaldehyde-assisted isolation of regulatory elements), which separates nucleosome-containing from nucleosome-free DNA using formaldehyde cross-linking followed by phenol extraction [57, 58]; or ATAC-seq [59], in which accessible chromatin is a preferential target for transposon tagging, allowing for direct sequencing of the tagged sequences after DNA isolation. ChIP-seq can also be used for genome-wide CRM discovery. For example, enhancer regions are often associated with the transcriptional cofactor p300/CBP and with components of the Mediator complex [60–62], and active enhancers are associated with specific histone modifications such as histone H3 lysine 4 monomethylation (H3K4me1) and histone H3 lysine 27 acetylation (H3K27ac), as well as depletion in H3 lysine 4 trimethylation (H3K4me3) [63, 62]. These methods all show great promise, although to date most have not yet produced detailed and well-defined sets of validated CRMs in the way that the traditional reporter gene assays (above) or newer functional assays (below) have done.

Function-Based Methods

The explosion in next-generation sequencing-based technologies has continued in the last few years with the development of new high-throughput function-based methods for enhancer discovery. STARR-seq can directly and quantitatively assess enhancer activity in millions of short sequences (on average ~600 bp in length) drawn from arbitrary sources of DNA to generate an unbiased survey of regulatory sequences active in a given cell line [64]. These sequences are inserted downstream of a minimal promoter and transfected into cells such that each sequence serves as its own reporter; the strength of each regulatory sequence is then assessed by its abundance in a subsequent RNA-seq analysis. When applied to the *Drosophila* genome, STARR-seq identified thousands of cell type-specific enhancers with differing activation strengths. Enhancer-FACS-seq is another method that was developed for identification of enhancers in *Drosophila*, where developmentally relevant, tissue-specific enhancers were detected within developing *Drosophila* embryos using a two-color FACS (fluorescently activated cell sorting)-based filtering: one color is used to register reporter gene activity and the other to mark cell types of interest [65]. This is an innovative method in that it eliminates the initial need to screen individual enhancer constructs in transgenic animals and allows for simultaneous testing of multiple pooled putative regulatory sequences, although full characterization of identified CRMs still requires subsequent generation of a new transgenic line. FIREWACh (Functional Identification of Regulatory Elements Within Active Chromatin) [66] and SIF-seq (site-specific integration fluorescence-activated cell sorting followed by sequencing) [67] also identify regulatory elements

by monitoring activity during initial screening assays using FACS sorting. Although they have not to date been applied to insect models, nothing specifically precludes their use for this purpose.

While these methods are elegant and high-throughput and demonstrate successful CRM discovery, they do have limitations. In particular, with respect to insect regulatory genomics, each depends either on the availability of a reasonable selection of cell lines or on the capacity to generate transgenic animals in an efficient and scalable manner—capabilities that for the most part are absent for insect species other than *D. melanogaster*.

Generality of Assays and Results

Although genome-wide maps of accessible chromatin, epigenetic marks, TF binding, and even regulatory function serve as a useful starting point, a significant challenge remains in that many regulatory regions function only in specific cell types and thus can only be identified when assays are performed using those cells. Each of these features must therefore be assessed in multiple tissues over many developmental time points and/or under varying environmental conditions in order to achieve comprehensive CRM discovery. This is a difficult goal for a variety of reasons, not least of which is financial, as well as obtaining sufficiently large homogeneous pools of each cell type at different time points and, more importantly, addressing depth of coverage in terms of the number of TFs and histone modifications to assay. These issues are especially acute in studying insects, which are anatomically small, thereby making it hard to isolate specific tissues in adequate amounts. In this regard, it is encouraging that DNaseI-seq at least appears to be reasonably robust in the sense that open chromatin regions are detected even when present in a limited fraction of overall embryonic cells [68]. Moreover, given the rate of technological progress, many of the logistical hurdles may soon be overcome as assays for small numbers of or even single cells are perfected [69, 70], and these methods will continue to aid in painting a more complete picture of the regulatory landscape of many cell types.

Assigning CRMs to Target Promoters

Once a CRM is identified, a major hurdle still often lies in assigning it to the appropriate target gene (or genes). Although many studies use “the closest active gene” theory to assign target genes, this clearly does not always lead to accurate assignment. Genes can lie hundreds of kilobases away from their cognate enhancers, and there can even be additional, separately regulated genes lying between a CRM-promoter pair. High-throughput versions of chromosome conformation capture technologies yield three-dimensional interaction maps that are providing exciting new insights into how distal CRMs interact with target promoters and can aid in CRM target gene assignment [30, 41, 71, 72], although these assays are technically challenging and artifact prone. Computational methods that make use of multiple

sources of more readily available data—histone modifications, RNA-seq, sequence conservation, etc.—will also be a valuable aid for determining CRM targets [72].

6.2.2.2 Computational Approaches to CRM Discovery

Even with the current trend of decreasing costs for empirical high-throughput experiments, the methods discussed in the preceding section remain prohibitively expensive and technically challenging for many emerging/non-model organisms, especially if considering extensive assaying of the genome under multiple conditions or at many developmental stages. Computational methods provide an attractive complement to experimental approaches and can often precede them as a first step in identifying regulatory regions, to be followed later by in vivo validation. Computational analysis can also help to refine or increase the predictive power of results obtained by empirical assays. In many cases, when working with non-model organisms with limited amenability to molecular genetic approaches, these methods may be essential for successful discovery and understanding of transcriptional regulatory elements.

Computational methods for CRM discovery can be broadly classified into three major categories: (a) comparative genomics, based on searching for regions of conserved noncoding DNA sequences across related species; (b) motif-based methods, which search for short genomic regions containing clusters of transcription factor binding sites; and (c) "motif-blind" approaches, which require no a priori knowledge of TFs or TFBSs.

Comparative Genomic Approaches

Comparative genomic approaches look for regions in the genome that are conserved between species. The underlying assumption is that there is likely to be a high degree of conservation of functionally important sequence elements (both coding and noncoding) between related species, an assumption that has frequently, although not universally, been shown to be true (e.g., [73] and references therein). There is mixed evidence as to whether or not attempting to discriminate CRMs from non-CRMs based solely on sequence conservation is effective. Li et al. [74] showed that while in the aggregate CRMs are more highly conserved, comparison among eight sequenced drosophilids gave poor predictive value for any particular sequence when assessing overall percentage of conserved bases. However, a more recent study found that reasonable discriminative performance could be achieved on a similar set of CRMs using a windowed version of the PhastCons conservation score (although on other data sets, this method performed less well) [75].

Less important than overall conservation of CRM sequence appears to be the conservation of CRM content, i.e., maintenance of a similar complement of TFBSs, although the number and organization of these sites can vary widely [76, 77]. As a result, sequence conservation is more clearly of utility when mixed with identification

of TFBSs, either to reduce false-positive identification of bona fide binding sites or to predict CRMs based on TFBS composition. For instance, the specificity of motif-based CRM prediction (see following section) can be improved by restricting TFBS motif instances to those that are also conserved in other species [78, 79]. Many in vivo-bound TFBS motifs are conserved among *Drosophila* species and other insect species [80–83]. Regulatory regions have been identified in several *Drosophila* genomes [80, 84, 85] as well as in other dipterans, including the malaria mosquito *An. gambiae*, the distant drosophilid *Scaptodrosophila lebanonensis*, and the fly *Calliphora vicina* [86–88], by looking for conservation of validated or predicted TFBSs when compared against the *D. melanogaster* genome. The enhancers from various species identified in this manner function as expected when tested in transgenic *Drosophila* [89–93]. Nevertheless, care must be taken when imputing function based on overall conservation of either sequence or binding site content. Studies of CRM evolution have revealed large-scale turnover of TFBSs despite the CRMs having maintained their function across multiple species of *Drosophila* [94, 95]. A landmark study by Ludwig et al. demonstrated that the *eve_stripe2* CRMs from *D. melanogaster* and *D. pseudoobscura*, which show clear sequence conservation as well as conservation of function, are completely nonfunctional as a chimera consisting of the 5′ half of one CRM and the 3′ half of the other [96]. Thus, TFBS turnover and compensatory evolutionary adaptations in the individual CRMs play a significant role in shaping their respective functions despite overall sequence-level conservation. Moreover, extensive enhancer mutagenesis has shown that simple scrambling of a CRM sequence can confer new tissue specificity to its output, and minor changes in motif positioning can affect CRM function in a tissue-specific manner [97, 98]. Merely possessing the same TFBSs, therefore, does not guarantee conservation of CRM function.

A significant limitation to sequence conservation as a means of CRM discovery, especially within the insects, is that noncoding sequences have evolved rapidly. Indeed, even within the Diptera, regulatory sequences have frequently diverged beyond the point of recognition by standard alignment methods ([96, 99–102] and M. Kazemian, S. Sinha, K.S and M.S.H., unpublished data). Moreover, sequence conservation cannot not reveal lineage-specific, recently evolved CRMs. Nevertheless, given the generality of the methods and the lack of need for any a priori knowledge of TFBS or TFs, comparative genomic approaches will remain a useful tool—at least for closely related species—for identification of putative CRMs.

Motif-Based CRM Discovery

In essence, CRMs are composed of a set of specific TFBSs spread over up to a few hundred nucleotides [103]. When these TFBSs are known or can be inferred, motif-based approaches for predicting enhancers and promoters can be applied. These approaches predict CRMs based on their DNA sequence and searchable representations of the TFBSs. Most typically, TFBSs are modeled in the form of position weight matrices (PWMs) [104], although alternate representations such as

degenerate consensus sequences or hidden Markov models have also been used [105, 106].

Motif-based CRM discovery was first conducted in mammals in the late 1990s in seminal work by Wasserman and Fickett [107]. Prior knowledge of the transcription factors that regulate expression of genes controlling muscle development, and their cognate TFBS motifs, was successfully used to look for clusters of those TFBS motifs elsewhere in the human genome. This approach was seized on by *Drosophila* researchers upon publication of the fly genome in early 2000 [108–113]. The extensive existing knowledge of early developmental CRMs—e.g., the "stripe" enhancers of the pair-rule genes—provided a rich set of TFBS motifs as well as a validation enhancer set to gauge sensitivity, and the tendency toward homotypic clustering of TFBSs within these CRMs allowed for simple "motif clustering" algorithms to be successful. All of these analyses found at least one novel enhancer that was active in transgenic flies, but in general suffered from low predictive power. A subsequent generation of algorithms incorporated probabilistic searching and sequence conservation between related species [114, 115], which helped to reduce false-positive rates; however, false-positive results continue to plague most motif-based CRM discovery methods, which are consistently outperformed in head-to-head comparisons of methods [75, 116, 117]. The high false-positive rates are likely a consequence of several factors, one of the largest being the fact that TFBS prediction itself is highly error-prone [118]. TFBS motifs are degenerate, and our knowledge of the full range of sequences capable of being bound by a given TF is usually incomplete, especially with respect to *in vivo* versus *in vitro* binding.

It is worth noting that motif-based methods rely on an important biological assumption: that genes that are expressed in a similar pattern are regulated by a similar complement of TFs. While no doubt this does not hold universally, the fact that these methods consistently work—albeit with high false-positive rates—supports the assumption. Nor is this confined only to the presence of highly tissue-specific TFs, broadly general expression patterns, or highly clustered binding sites. For instance, Halfon et al. [112] demonstrated that motif-based searching could identify CRMs driving a tightly restricted expression pattern in a small subset of cells and regulated by a combination of widely expressed TFs, some of which bound only once or twice in the CRM. On the other hand, while it is clear that identification of (usually conserved) TFBSs can aid in CRM discovery, caution must be taken in ascribing functional roles to each of these sites and/or to their cognate TFs. Previous studies have shown that not all motifs used as input for successful CRM discovery algorithms are functional components of the identified CRMs, and not all important TFBSs are conserved [119, 120].

A different flavor of CRM discovery moves away from clustering of a specific set of TFBSs toward a model of CRM evolution via gain and loss of binding sites. These methods attempt to develop mathematical models that capture the TFBS signatures characteristic of CRMs without assuming direct sequence-level conservation. MorphMS is one such modeling method which identifies candidate CRMs using a pairwise probabilistic alignment method that fits an evolutionary model derived from a set of existing TFBS motifs; it was found to have the best performance

for recovering known *D. melanogaster* CRMs in a comparison of computational approaches [75, 121]. EMMA, an improvement on MorphMS by the same authors, models the evolution of binding sites and allows binding sites to occur in only one species, but not the other (note that both tools construct pairwise alignments) [122]. A similar approach, to account for gain and loss of binding sites, is taken by Majoros and Ohler [123], although its computational complexity precludes it from being implemented on a genome-wide scale. These approaches provide important insights into the potential roles of TFBS turnover in CRM evolution. However, as CRM discovery methods, they still suffer from the requirement of needing knowledge of relevant TFBSs to be effective.

Motif-Blind Approaches

What does one do, then, when not all TFs and/or TFBSs (or sometimes not even one relevant TF and/or its TFBS) are known a priori—the most common situation? In such an event, it becomes impossible to search for CRMs using motif-based methods, and it is necessary to turn to methods that are not limited by current knowledge of TFBS motifs or of the TFs involved in regulating genes of interest. This becomes especially crucial for annotating the genomes of non-model organisms (such as most insect species) where such data are severely lacking.

One approach that has been used is to employ motif discovery and CRM discovery in tandem. An example of this is CisModule, which uses a Bayesian model to simultaneously predict TFBS motifs and CRMs [124]. CisModule showed good specificity in both simulated and applied tests, particularly for its motif-finding phase. However, in other settings, it performed less well for de novo CRM discovery than motif-blind (see below) methods that do not rely on first predicting TFBSs [116].

Better success has been achieved using supervised machine learning methods which search for patterns that can distinguish a training set composed of known CRMs from non-regulatory DNA, using only the DNA sequence itself as input [117, 125, 126]. These methods capture the statistical features inherent in each CRM within the training set without requiring other information, such as TFBSs or TFs, a priori. The genome can then be searched for additional sequence windows containing a similar statistical signature. Kantorovitz et al. [117] first dubbed such methods, which fall into the class of alignment-free sequence comparisons, "motif-blind," as TFBS motifs do not factor into the search algorithms.

One of the most successful examples of motif-blind approaches has come from a collaborative effort between the Sinha and Halfon groups, who in a series of papers have applied their methods to both the *Drosophila* and mouse genomes [117, 126]. This team has developed a computational pipeline, designated "SCRMshaw," that uses multiple machine learning algorithms to search for sequence "words" (i.e., short DNA subsequences) that are overrepresented in a training set of known CRMs (Fig. 6.3). These words (or "*k*-mers") serve as proxies for the unknown and un-modeled TFBSs, but TFBSs themselves, even when known, are not explicitly used by the algorithms. The training sets are constructed from a set of CRMs all

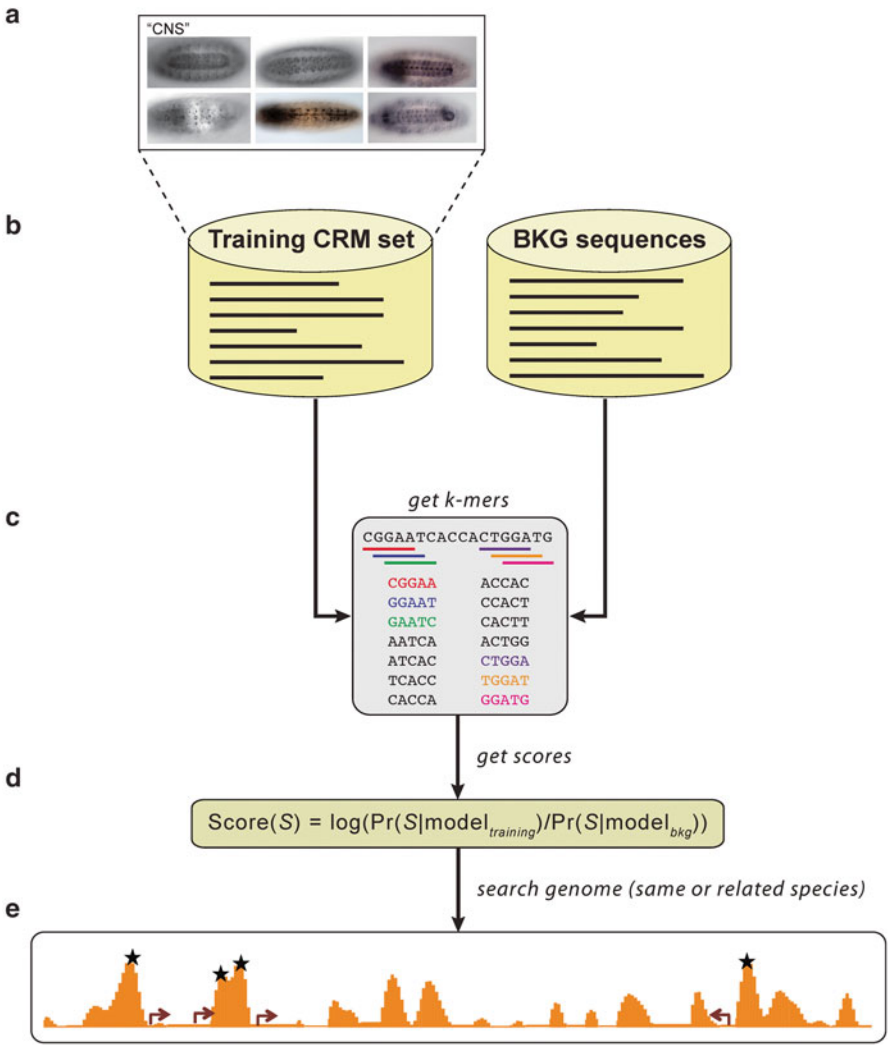


Fig. 6.3 Supervised motif-blind CRM discovery (SCRMshaw). (**a**) A set of CRMs with related activity (shown here, the *Drosophila* embryonic ventral nerve cord, part of the "CNS" CRM set) is selected as a training set. (**b**) The sequences of the training CRMs and of a set of similarly sized "background" non-CRMs (BKG) serve as input to the algorithm. The training set can also include orthologous sequences from related species. (**c**) The k-mer profile of the sequence sets is obtained and used to train one of the several statistical models. (**d**) The score for a given sequence S is the log-likelihood ratio of the models for the positive ("training") and negative ("background") sets on S. (**e**) Overlapping sequence windows are scored throughout the genome. High-scoring windows (*stars*) are predicted CRMs. The genome being searched can be from the same species as the training data (e.g., *Drosophila melanogaster*) or from a more distantly related insect species (e.g., *Apis mellifera*)

demonstrated to drive a related expression pattern and do not need to be large; sets as small as six known CRMs have provided successful CRM discovery results. Comparisons with motif-based methods for CRM sets where there is good knowledge of TFBSs—e.g., the *Drosophila* stripe enhancers referred to above—demonstrated that SCRMshaw consistently performed as good or better and was able to reach unprecedentedly high (80 % or better) success rates [117].

CRM Discovery in Insects Other Than *D. melanogaster*

The number of characterized non-*Drosophila* insect CRMs is small but growing. Computational methods based on motif clustering have proven effective in discovering CRMs in insects other than *Drosophila*, usually using PWMs derived from *Drosophila* TF binding studies and relying on the assumption that the binding sites for orthologous TFs would have similar sequences [101, 127–130]. These studies have mainly focused on well-described developmental systems, in particular early anterior–posterior and dorsal–ventral patterning, where there is extensive knowledge of TFs, their binding motifs, and similar *Drosophila* CRMs.

We have recently determined that *Drosophila* CRM training data can be used to apply the SCRMshaw method for motif-blind supervised CRM discovery to a broad range of holometabolous insects with success rates comparable to those obtained when conducting *Drosophila*-specific CRM discovery [131]. By using the same methods and training sets used for within-species CRM discovery [117, 126] but searching the genomes of *An. gambiae*, *T. castaneum*, *A. mellifera*, and *N. vitripennis* instead of that of *D. melanogaster* (Fig. 6.3), we were able to rapidly almost double the collective number of in vivo validated CRMs for these species and predict some 7000 more [131]. This is a significant advance given that the genomes of these species are highly diverged—substantially more so than human-to-fish for Diptera-to-Hymenoptera, for example [132]—to the point where alignment of noncoding sequences to the *Drosophila* genome is for the most part not possible. Successful application of supervised motif-blind CRM discovery therefore suggests that not only is regulatory sequence annotation in diverged insect species an attainable goal but also that it is one that can progress without requiring extensive new experimental data to be generated for each newly sequenced genome.

6.2.2.3 Database Resources for Insect Genomic Data

Biological databases are an essential part of any research project undertaken today. The ever-increasing amounts of data collected from biological experiments, especially high-throughput experiments such as genome sequencing and annotation, protein and gene interaction studies, protein structure determination, and the like, make these databases invaluable for managing information and making it easily accessible. Several dedicated databases have been developed for insect-specific research and are briefly reviewed below.

Model Organism Databases

Many of the insects that have been sequenced within the last decade now have dedicated model organism databases (MODs) [133–141]. The MODs constitute a tremendously valuable resource and serve as clearinghouses for much of the available data on the genetics and genomics of the covered model organisms. This is most evident for *Drosophila*, where FlyBase, one of the first MODs to be developed, maintains not just the genome annotation but also allele descriptions, gene expression pattern data, transcriptomic data, cytogenetic maps, and much of the other collected information from over a century of *Drosophila* research [133].

While the MODs are crucial for allowing researchers to access sequence data and genome annotations, a problem often encountered with species-specific databases is that of interoperability. Different interfaces and data formats make it complicated for users to move about through the different databases, and the databases include widely varying degrees of information on homologous sequences in other species, tools for pathway analysis, gene ontology annotations, protein domain annotation (e.g., InterPro), and functional pathway annotation (e.g., KEGG). In this regard, the Hymenoptera Genome Database stands out as a truly multispecies genome database for representatives of the over 115,000 insects in the Hymenopteran order [134]. Combining information on all these species into a single database provides an enormously useful resource for researchers interested in comparing and studying the Hymenoptera. The combination of numerous pest species into the AgripestBase (www.agripestbase.org) framework is another positive step in the direction of interoperability. As various species are becoming sequenced through the i5K project [2], many of the assemblies and early annotation are being housed through the National Agricultural Library's "i5K Workspace" (http://i5k.nal.usda.gov/), which provides a hosting framework for species not backed by a large, organized research community. The i5K Workspace and many of the MODs are built using components from the Generic Model Organism Database (GMOD) toolkit [142], a powerful resource for researchers who wish to provide bioinformatic tools for accessing whole-genome data. The use of GMOD components by a broad selection of MODs provides a familiar interface and a degree of interoperability for users of multiple genome databases. A holdout in this regard is VectorBase [143], which is built using the ENSEMBL framework rather than GMOD. Although this design choice has many positive features—the versatile and intuitive BioMart [144] is a particularly useful tool—it places VectorBase somewhat at odds with the other MODs and complicates cross-organism comparisons. Some of the more traditional model insect species (several *Drosophila*, *An. gambiae*, *A. mellifera*) can also be found in the UCSC Genome Browser [145], allowing access to the powerful tools, and integration with the many other genomes, covered by that major resource. Similarly, many insect species are also accessible via ENSEMBL (http://metazoa.ensembl.org/index.html). *Drosophila* and *An. gambiae* data can be found in FlyMine, a data warehouse with a powerful search interface that integrates genomic and proteomic data for these two species [146]. While these latter three databases offer the advantages of data integration and standard included tools, it should be

noted that the primary genome sequences and annotations are still imported from the MODs.

Gene Expression Resources

Several resources are devoted to gene expression data. The Berkeley *Drosophila* Genome Project (BDGP) contains genome-wide expression profiles of over 6000 genes in *D. melanogaster* embryos as determined by whole-mount in situ hybridization over all embryonic developmental stages and documented in over 70,000 images [147–149]. FlyExpress is another such resource that catalogues the spatial expression domains of over 4000 genes via a series of over 100,000 images and allows for pattern-based searching of the database [150]. FlyAtlas [151] provides transcriptional profiles for dissected *D. melanogaster* adult and larval tissues, and modENCODE [152] has produced time-course expression data for all stages of the fly life cycle as well as a limited number of dissected tissues. Many of these data are also mirrored in FlyBase. Many of the other MODs also include gene expression data, either from EST sequencing, microarray, or RNA-seq studies (see Table 6.1).

6.2.2.4 Regulatory DNA Element and Transcription Factor Databases

Resources related to insect gene regulation are primarily directed toward *Drosophila*, where the bulk of the existing work on regulatory element discovery has been performed. The most comprehensive regulatory genomics database available for insects—in fact, for any metazoan—is REDfly, the Regulatory Element Database for *Drosophila* [153]. REDfly is a highly curated portal for *Drosophila cis-regulatory* data containing records for empirically validated CRMs and TFBSs obtained from the published literature. This single searchable database of CRMs enables researchers to search for all experimentally verified fly regulatory elements along with their DNA sequence, their associated genes, and the expression patterns they direct (Fig. 6.4). REDfly serves as an important source of data for both validation and generation of hypotheses about gene regulation and has been particularly important for facilitating studies of CRM evolution and development of methods for CRM discovery.

The JASPAR [154] and TRANSFAC [155] databases are a major source of TFBS data, but although they contain TFBSs from *Drosophila*, they are not limited to insects, and most of their data are from vertebrate species. FlyFactorSurvey, on the other hand, contains *D. melanogaster* TF binding specificities as determined by bacterial one-hybrid assays, SELEX, or DNase I footprinting. The database contains PWMs associated with over 300 TFs and computational tools for identifying motifs within new candidate sequences [156]. The related Genome Surveyor [157] is a web-based tool for CRM discovery and analysis in a growing number of species; covered insect species include *D. melanogaster*, *Ae. aegypti*, *An. gambiae*, *N. vitripennis*, *A. mellifera*, and *T. castaneum*. Using the motifs contained within

Table 6.1 Database resources for insect regulatory genomics

Resources	Description	Species included (as of June 2015)	Link
AgripestBase	A comprehensive model organism database for agricultural pests	The Hessian fly, *Mayetiola destructor*; the tobacco hornworm, *Manduca sexta*; and the red flour beetle, *Tribolium castaneum*	http://agripestbase.org/
AphidBase	Model organism database	The pea aphid *Acyrthosiphon pisum*	http://www.aphidbase.com/
Berkeley Drosophila Genome Project in situ database	Contains genome-wide spatial expression profiles of 7917 genes during embryogenesis	*Drosophila melanogaster*	http://insitu.fruitfly.org/
Ensembl–Metazoa	A database for genomes of metazoan species including a number of insect species, with tools for querying and extracting features of each genome such as sequence variation, annotation, and protein homologies	Genomes of 17 dipteran, 4 hymenopteran, 4 lepidopteran, 2 coleopteran, 2 hemipteran, 1 isopteran, and 1 pthirapteran species	http://metazoa.ensembl.org/index.html
FlyAtlas 2	Transcriptional profiles of genes in multiple tissues at multiple larval through adult developmental stages	*D. melanogaster*	http://flyatlas.gla.ac.uk/flyatlas/index.html
FlyBase	Model organism database	Twelve species in the genus *Drosophila*	http://flybase.org/
FlyExpress	Digital library capturing the spatiotemporal expression patterns of thousands of genes during development. Can be used to match/search for specific expression patterns of interest	*D. melanogaster*	http://www.flyexpress.net/
FlyMine	An integrated resource for multiple types of genomic and proteomic data for *Drosophila* and *Anopheles*	*D. melanogaster*, *An. gambiae*	http://www.flymine.org/
FlyTF	An integrated database of data for *Drosophila* transcription factors	*D. melanogaster*	http://www.flytf.org/

(continued)

Table 6.1 (continued)

Resources	Description	Species included (as of June 2015)	Link
Genome Surveyor	A web-based tool for discovery and analysis of *cis-regulatory* elements in *Drosophila* and other organisms. Provides prediction and visualization of putative CRMs and TFBSs	*D. melanogaster*, *An. gambiae*, *A. mellifera*, *N. vitripennis*, *T. castaneum*	http://veda.cs.uiuc.edu/cgi-bin/gb2/gbrowse/Dmel5/
Hymenoptera Genome Database	Model organism database	Species in the order Hymenoptera including 3 *Nasonia* species, 4 bee species, and 8 ant species; also the genomes of 4 *Apis mellifera* pests and pathogens	http://hymenopteragenome.org
JASPAR	An excellent resource for a curated, nonredundant set of TF binding profiles, derived from published collections of experimentally defined transcription factor binding sites for several organisms; tools for querying DNA sequences of interest for instances of TFBS in the database	*D. melanogaster*; various noninsects	http://jaspar.genereg.net/
LocustDB	A transcriptomic database with a library of ESTs	The migratory locust *Locusta migratoria*	http://locustdb.genomics.org.cn/
modENCODE	Data access portal for genomic and epigenomic data from the modENCODE project	*D. melanogaster* and several other *Drosophila* species; *Caenorhabditis* species	http://www.modencode.org/
MyzusDB	A preliminary database resource with whole genome as well as comparative genome analyses	The green peach aphid *Myzus persicae*	http://www.aphidbase.com/node_94263/Myzus-DB

(continued)

Table 6.1 (continued)

Resources	Description	Species included (as of June 2015)	Link
ORegAnno	An open database that allows users to manually curate and annotate regulatory elements as well as visualize and access the annotated regulatory elements	*D. melanogaster* but mostly noninsect species	http://www.oreganno.org/oregano/
REDfly	Comprehensive database of over 5000 experimentally validated CRMs and TFBSs along with accompanying information such as expression patterns, sequences, and target genes	*D. melanogaster*	http://redfly.ccr.buffalo.edu/
SilkDB	Model organism database	*Bombyx mori*	http://www.silkdb.org/silkdb/
SpodoBase	Model organism database	*Spodoptera frugiperda* (fall army worm)	http://bioweb.ensam.inra.fr/spodobase/
TRANSFAC	A manually curated database of eukaryotic transcription factors, their genomic binding sites, and DNA binding profiles	*D. melanogaster*; many noninsects	http://www.gene-regulation.com/pub/databases.html
UCSC Genome Browser	A comprehensive resource for all genomic information as well as proteomic information for several model and non-model insect species	11 drosophilids, *An. gambiae*, and *A. mellifera*; many noninsects	https://genome.ucsc.edu/
VectorBase	Model organism database	19 *Anopheline* species, *Ae. aegypti*, *C. quinquefasciatus*, the Tsetse fly *G. morsitans*, many others	https://www.vectorbase.org/

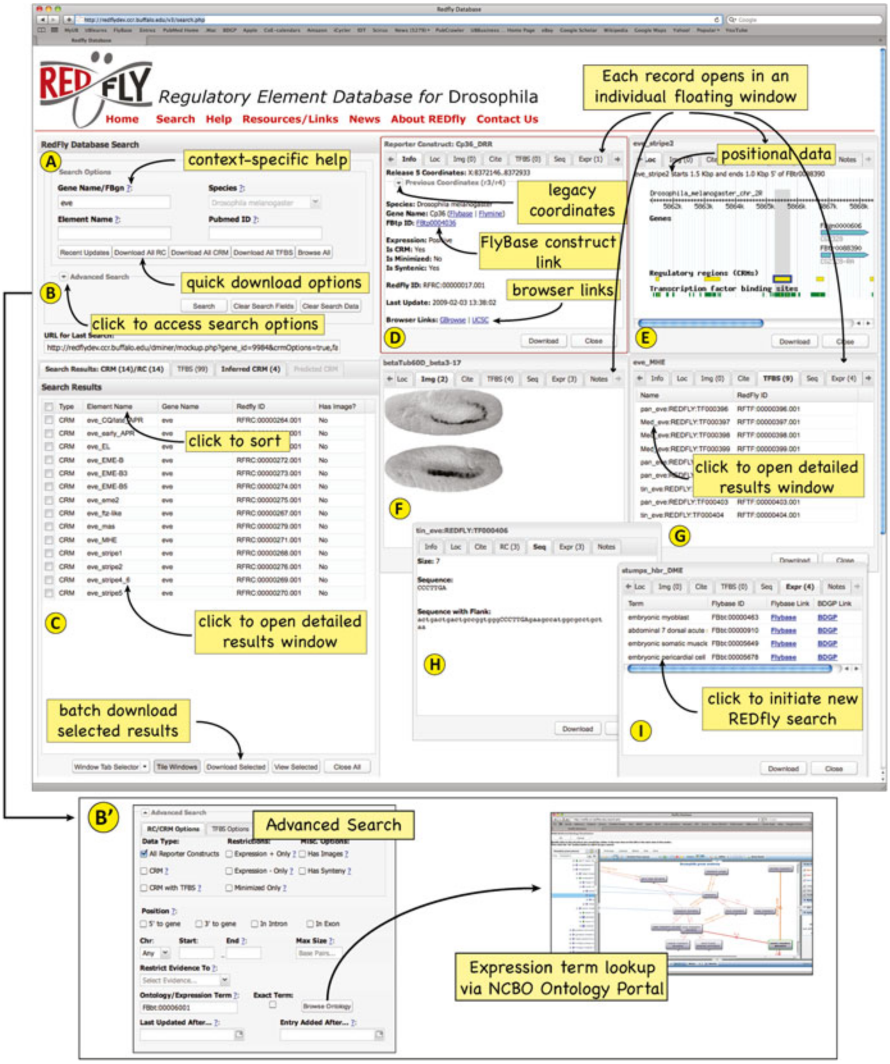


Fig. 6.4 The REDfly database REDfly is a comprehensive CRM and TFBS database for *Drosophila*. Search options (*A*, *B*), results overview (*C*), and detailed results (D–I) are all displayed within a single web browser window. Advanced search options (*B'*) include the ability to search based on ability of a tested genomic sequence to regulate/not regulate gene expression, position of a CRM relative to the transcription start site of the gene, and pattern of expression regulated by the CRM. For the latter function, an anatomy ontology browser can be used to select desired search terms (*right-hand panel*). The detailed results (*D–I*) are displayed as individual floating windows that can be stacked or tiled to facilitate comparison of multiple CRMs (Adapted from Gallo et al. [153])

FlyFactorSurvey, Genome Surveyor can predict TFBSs and CRMs in *Drosophila* using several different methods, including the supervised motif-blind CRM discovery method from Kantorovitz et al. [117]. FlyTF [158] allows for query-based retrieval of curated TFs for several *Drosophila* species that have been identified using different biological assays such as footprinting and chromatin interaction assays, and their target genes, although at present it is no longer being actively maintained. A newer resource for *D. melanogaster* TFs, OnTheFly, includes TFs, their binding sites, and annotation of their DNA-binding domains with structural properties and evolutionary homology [159].

The rapid accumulation of experimental data in the field of insect genomics highlights the need for databases that include interactive web-based computational analysis tools to simplify integration of different types of data such as genome-wide high-throughput genomic data, proteomic data, transcriptome data, and RNAi data with genome annotations for transcripts and regulatory elements. FlyMine does much of this for the two species it covers, but does not currently provide a home for additional insect species. REDfly would be a natural repository for regulatory-specific data from across the Insecta as they become available and was designed with this goal in mind, although to date no non-*Drosophila* data have been incorporated. Galaxy [160–162] provides a user-friendly platform for conducting many types of genomic analysis and has potential as a unifying tool for bringing together different data sources [163]. Although having a single consolidated resource for insect genomics and proteomics would greatly facilitate research and reduce the need to navigate multiple different database implementations, developing tools and methods to better take advantage of existing resources for data integration and analysis may prove the most feasible and cost-effective strategy.

6.3 Insect Transgenesis: Historical Perspective and Current State

6.3.1 Application to Understanding Gene Regulation

The ability to transform foreign DNA into a host genome has proven to be a powerful tool for genetic analysis and manipulation and is instrumental for studies of gene regulation. Transgenesis allows for in vivo reporter gene analysis, essential for characterizing regulatory sequences, as well as for generating cell- and tissue-specific markers, determining cell lineages, ablating specific cells, and marking chromosomes for genetic studies.

Genetic transformation was first applied to insects almost half a century ago in the flour moth *Ephestia kuehniella* [164]. In this experiment, larvae with mutant wing scales were injected with wild-type DNA, with some developing into adults with rescue of the phenotype from integrated DNA. Microinjection of DNA into embryos began in the late 1970s with mutant rescue experiments in *D. melanogaster* [165], but *Drosophila* transgenesis did not really take off until the seminal development

by Rubin and Spradling of stable germline transformation through the use of the *P* element transposon [166, 167]. This marked the first instance of mutant rescue in an animal model by heritable gene transfer and laid the foundation for germline transformation in many other model organisms. Although *P*-based transformation proved ineffective for other insect species, a number of other transposable elements with broad efficacy in insects have since been identified. *Minos*, originally discovered in *D. hydei*, was the first transposon vector that was successful in transformation of a non-drosophilid, the medfly *Ceratitis capitata* [168]. The *Minos* transposon is especially useful because of its low insertional bias and high-frequency transformation rates and has thus seen wide use in vertebrate and invertebrate model organisms alike [169]. A second transposon, *piggyBac*, discovered in the cabbage looper moth *Trichoplusia ni* [170], is perhaps the most widely used transposon vector to date and has seen use for transgenesis in many eukaryotic systems [171], including insects [172] and even human cells [173]. *piggyBac* has been used to extend enhancer trapping strategies for identification and functional analysis of genes in both *B. mori* and *T. castaneum* [9, 174–179] and has also been used to transform the genomes of the butterfly *Bicyclus anynana* and the honey bee *A. mellifera* [180, 181]. *Bicyclus* has also been transformed with the transposon *Hermes* [180].

More recently, the use of site-specific recombinases has allowed for reproducible insertion into specific loci. One of the most highly used systems is the φC31 integrase [182–187]. A high and stable integration frequency coupled with its ability to accept integration of large inserts (over 100 kb) has made this a method of choice for many *Drosophila* applications. Subsequent reports have demonstrated the utility of this integrase system in *Ae. aegypti*, *Ae. albopictus*, *An. gambiae*, *C. capitata*, and *B. mori* [188–194]. Although φC31 integration shows great promise for facilitating efficient transformation in diverse insect species, an important caveat is that its use requires prior engineering of the host genome to insert an *attP* landing site for the integration event; multiple landing site choices are desirable as not all sites may prove effective for all applications. Therefore, φC31-mediated transgenesis has been restricted to species for which at least one other method for germline transformation is already available, so that landing site strains can be constructed. However, the relative ease of CRISPR-/Cas9-based genome engineering may soon make it possible to readily add integration landing sites or to simply insert transgenes at a desired location, in a species of choice. Indeed, while the genomes of most insects historically have been refractory to manipulation, the i5K project has provided an impetus to develop and apply efficient transgenic technology to better take advantage of the wealth of accumulating sequence data, and it is likely that we will soon see rapid improvements in strategies for insect transgenesis.

6.3.2 Biotechnological Applications

Effective insect transgenesis will be instrumental to further studies of insect biology and to the understanding of insect gene regulation, and the ability to combine transgenic technologies with a firm understanding of regulatory genomics carries

exciting potential for developing improved methods for insect management and control. For example, the ability to drive gene expression in the adult female salivary glands, midgut, and fat body of *Anopheline* mosquitoes (tissues that play critical roles during infection by and transmission of malaria-causing *Plasmodium* parasites) should greatly facilitate studies of mosquito/parasite interactions and may eventually lead to improved strategies for malaria mitigation [195]. A CRM of the *nanos* gene has been used to drive gene expression in female germ cells of *Ae. aegypti*, a key innovation in mosquito transgenic technology with major implications for future genetic engineering and improved population control of this important disease vector [196]. Similarly, the recent development of female-flightless transgenic control strategies for *Ae. aegypti* uses a muscle-specific CRM to ablate flight muscles in adult females, leading to flightless and therefore effectively sterile mosquitoes [194, 197]. In a materials science rather than a disease vector control setting, application of transgenic technology has been used in the silkworm *B. mori* to produce a variety of biomaterials including expression of the spider silk protein *MaSp1* driven by the *B. mori Ser1* promoter, resulting in silkworm-produced silk with the same unparalleled tensile and structural properties as spider dragline silk [198, 199] (see volume 2, Chap. 9, in this series). Elucidation of additional species-specific and tissue-specific regulatory elements, coupled with improved ability to construct transgenic insects, promises many more advances along these lines in the years to come.

6.4 Prospects for Studying Evolution

Changes in CRMs alter the structure and function of gene regulatory networks, making CRM evolution a major driving force of the morphological diversity seen in metazoan body plans [200–203]. New regulatory functions may be acquired not just by changes in existing CRMs [204–208] but also by the gain of entirely new enhancers, which can arise de novo from nucleotide substitution, deletion, insertion, transposition, or duplication. The details of these processes, as well as the relative frequency of CRM repurposing versus CRM creation, are not yet well understood. Insects are an ideal class of animals in which to study regulatory evolution due to their tremendous diversity and the growing number of species becoming amenable to experimental manipulation. The incredible morphological specialization found within insects even at the family and subfamily level provides us with the opportunity to build a comprehensive comparative developmental framework and to elucidate the genetic and molecular mechanisms behind the vast insect radiation.

6.5 Concluding Remarks

The past decade has witnessed dramatic progress in the area of regulatory genomics, driven by developments in genome sequencing and analysis. Insects, spearheaded by the model research animal *D. melanogaster*, have played a major role in these

advances. The next several years will see completion of the full genome sequencing of a large number of insects across a wide evolutionary spectrum. A great challenge will thus be annotating the regulatory genomes of these diverse sequenced species. Fortunately, the outlook is bright for non-model insects. Decreasing costs for genomic assays and the ability to apply them to increasingly small numbers of—or even single—cells [59, 69, 70, 209] raise the hope that direct empirical studies will become feasible for many different species. Similarly, methods such as RNAi and CRISPR-/Cas9-based genome engineering open up traditionally nongenetic systems to experimental analysis. Computational methods, which have matured greatly over the last dozen years, can predict with growing accuracy CRMs in model and non-model organisms alike. It is thus with great anticipation that we look forward to seeing the power of the computational and empirical methods developed for studying regulatory genomics applied broadly to the insects, with their enormous diversity and tremendous impact on human health and agriculture.

6.6 Further Reading

For a general treatment of transcriptional regulation in eukaryotes, see the detailed review by Maston et al. [33].

For a current perspective on enhancer biology and the implications of the myriad studies on the role of enhancers in development, disease, and evolution, see the recent set of commentaries by several prominent researchers in *Nature Genetics* [210].

For more on how TFBSs can be represented and the basis for such representations, see [104]. The review by Stormo [104] remains one of the best gentle introductions to PWM-based TFBS representation. For a brief and accessible yet thorough treatment, see [211].

For more on computational tools available for motif discovery, readers are referred to the excellent reviews by Zambelli et al. [212], a commentary on the different methods for TFBS discovery before and after the advent of next-generation sequencing, and MacIsaac et al. [213], which describes strategies for using motif-based methods and tools.

Reviews on methods for CRM discovery (both empirical and computational) include Haeussler and Joly's review on strategies and methods to choose when embarking on a CRM discovery project [214] and overviews of the many computational methods for CRM discovery by Van Loo and Marynen [215] and Aerts [216].

For a review on the numerous next-generation technologies currently available to aid functional genomics studies, readers are referred to the excellent commentary by Wold and Myers [217]. Zentner et al. [218] discuss using chromatin features to identify enhancers, and Shyueva et al. [219] provide a recent review on current technologies available for large-scale annotation of regulatory elements.

Resources for *Drosophila*-specific genomics are comprehensively reviewed by Mohr et al. [220].

For reviews on the role on enhancers and CRMs in evolution, see [73, 200, 203, 221]. For thorough coverage of this subject, two major treatments are the books by Eric Davidson [103] and Sean Carroll [222].

Acknowledgments The authors are grateful to John Nyquist at the University at Buffalo for help with the illustrations contained in this chapter. The authors are supported by USDA grant 2011-04656.

References

1. Banerji J, Rusconi S, Schaffner W (1981) Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27(2 Pt 1):299–308
2. i5K Consortium (2013) The i5K initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104(5):595–600
3. Liu J, Li C, Yu Z et al (2012) Efficient and specific modifications of the *Drosophila* genome by means of an easy TALEN strategy. J Genet Genomics 39(5):209–215
4. Gilbert LA, Larson MH, Morsut L et al (2013) CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. Cell 154(2):442–451
5. Elick TA, Bauser CA, Fraser MJ (1996) Excision of the piggyBac transposable element *in vitro* is a precise event that is enhanced by the expression of its encoded transposase. Genetica 98(1):33–41
6. Sarkar A, Coates CJ, Whyard S et al (1997) The Hermes element from *Musca domestica* can transpose in four families of cyclorrhaphan flies. Genetica 99(1):15–29
7. Berghammer AJ, Klingler M, Wimmer EA (1999) A universal marker for transgenic insects. Nature 402(6760):370–371
8. Peloquin JJ, Thibault ST, Staten R et al (2000) Germ-line transformation of pink bollworm (Lepidoptera: Gelechiidae) mediated by the piggyBac transposable element. Insect Mol Biol 9(3):323–333
9. Tamura T, Thibert C, Royer C et al (2000) Germline transformation of the silkworm *Bombyx mori* L. using a piggyBac transposon-derived vector. Nat Biotechnol 18(1):81–84
10. Pavlopoulos A, Berghammer AJ, Averof M et al (2004) Efficient transformation of the beetle *Tribolium castaneum* using the Minos transposable element: quantitative and qualitative analysis of genomic integration events. Genetics 167(2):737–746
11. Nakamura T, Yoshizaki M, Ogawa S et al (2010) Imaging of transgenic cricket embryos reveals cell movements consistent with a syncytial patterning mechanism. Curr Biol 20(18):1641–1647
12. Warren IA, Fowler K, Smith H (2010) Germline transformation of the stalk-eyed fly, *Teleopsis dalmanni*. BMC Mol Biol 11:86
13. Kennerdell JR, Carthew RW (2000) Heritable gene silencing in *Drosophila* using double-stranded RNA. Nat Biotechnol 18(8):896–898
14. Brown S, Holtzman S, Kaufman T et al (1999) Characterization of the *Tribolium Deformed* ortholog and its ability to directly regulate *Deformed* target genes in the rescue of a *Drosophila Deformed* null mutant. Dev Genes Evol 209(7):389–398
15. Terenius O, Papanicolaou A, Garbutt JS et al (2011) RNA interference in Lepidoptera: an overview of successful and unsuccessful studies and implications for experimental design. J Insect Physiol 57(2):231–245
16. Blandin S, Moita LF, Kocher T et al (2002) Reverse genetics in the mosquito *Anopheles gambiae*: targeted disruption of the *Defensin* gene. EMBO Rep 3(9):852–856
17. Osta MA, Christophides GK, Kafatos FC (2004) Effects of mosquito genes on Plasmodium development. Science 303(5666):2030–2032

18. Bucher G, Scholten J, Klingler M (2002) Parental RNAi in *Tribolium* (Coleoptera). Curr Biol 12(3):R85–R86
19. Lynch JA, Desplan C (2006) A method for parental RNA interference in the wasp *Nasonia vitripennis*. Nat Protoc 1(1):486–494
20. Lynch JA, Roth S (2011) The evolution of dorsal-ventral patterning mechanisms in insects. Genes Dev 25(2):107–118
21. Hughes CL, Kaufman TC (2000) RNAi analysis of Deformed, proboscipedia and Sex combs reduced in the milkweed bug *Oncopeltus fasciatus:* novel roles for Hox genes in the hemipteran head. Development 127(17):3683–3694
22. Belles X (2010) Beyond *Drosophila*: RNAi in vivo and functional genomics in insects. Annu Rev Entomol 55:111–128
23. Bibikova M, Golic M, Golic KG et al (2002) Targeted chromosomal cleavage and mutagenesis in *Drosophila* using zinc-finger nucleases. Genetics 161(3):1169–1175
24. Bassett AR, Tibbit C, Ponting CP et al (2013) Highly efficient targeted mutagenesis of Drosophila with the CRISPR/Cas9 system. Cell Rep 4(1):220–228
25. Gratz SJ, Cummings AM, Nguyen JN et al (2013) Genome engineering of *Drosophila* with the CRISPR RNA-guided *Cas9* nuclease. Genetics 194(4):1029–1035
26. Richter H, Randau L, Plagens A (2013) Exploiting CRISPR/Cas: interference mechanisms and applications. Int J Mol Sci 14(7):14518–14531
27. Bassett AR, Liu JL (2014) CRISPR/Cas9 and genome editing in *Drosophila*. J Genet Genomics 41(1):7–19
28. Yu Z, Ren M, Wang Z et al (2013) Highly efficient genome modifications mediated by CRISPR/Cas9 in *Drosophila*. Genetics 195(1):289–291
29. Ma S, Chang J, Wang X et al (2014) CRISPR/Cas9 mediated multiplex genome editing and heritable mutagenesis of BmKu70 in *Bombyx mori*. Sci Rep 4:4489
30. Dostie J, Richmond TA, Arnaout RA et al (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. Genome Res 16(10):1299–1309
31. Dekker J, Rippe K, Dekker M et al (2002) Capturing chromosome conformation. Science 295(5558):1306–1311
32. Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. Nat Rev Genet 13(4):233–245
33. Maston GA, Evans SK, Green MR (2006) Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 7:29–59
34. Kutach AK, Kadonaga JT (2000) The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. Mol Cell Biol 20(13):4754–4764
35. Zhu Q, Halfon MS (2009) Complex organizational structure of the genome revealed by genome-wide analysis of single and alternative promoters in *Drosophila melanogaster*. BMC Genomics 10:9
36. Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100(26):15776–15781
37. Ni T, Corcoran DL, Rach EA et al (2010) A paired-end sequencing strategy to map the complex landscape of transcription initiation. Nat Methods 7(7):521–527
38. Batut P, Gingeras TR (2013) RAMPAGE: promoter activity profiling by paired-end sequencing of 5′-complete cDNAs. Curr Protoc Mol Biol 104:Unit 25B 11
39. Mardis ER (2007) ChIP-seq: welcome to the new frontier. Nat Methods 4(8):613–614
40. Collas P, Dahl JA (2008) Chop it, ChIP it, check it: the current status of chromatin immunoprecipitation. Front Biosci 13:929–943
41. Fullwood MJ, Ruan Y (2009) ChIP-based methods for the identification of long-range chromatin interactions. J Cell Biochem 107(1):30–39
42. Pillai S, Chellappan SP (2009) ChIP on chip assays: genome-wide analysis of transcription factor binding and histone modifications. Methods Mol Biol 523:341–366

43. Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science 327(5963):335–338
44. Hoskins RA, Landolin JM, Brown JB et al (2011) Genome-wide analysis of promoter architecture in Drosophila melanogaster. Genome Res 21(2):182–192
45. Kvon EZ, Kazmar T, Stampfel G et al (2014) Genome-scale functional characterization of Drosophila developmental enhancers in vivo. Nature 512(7512):91–95
46. Jory A, Estella C, Giorgianni MW et al (2012) A survey of 6,300 genomic fragments for *cis-regulatory* activity in the imaginal discs of *Drosophila melanogaster*. Cell Rep 2(4):1014–1024
47. Jolma A, Kivioja T, Toivonen J et al (2010) Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res 20(6):861–873
48. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc 4(3):393–411
49. Meng X, Brodsky MH, Wolfe SA (2005) A bacterial one-hybrid system for determining the DNA-binding specificity of transcription factors. Nat Biotechnol 23(8):988–994
50. Rhee HS, Pugh BF (2012) ChIP-exo method for identifying genomic location of DNA-binding proteins with near-single-nucleotide accuracy. Curr Protoc Mol Biol 100:21.24:21.24.1–21.24.14
51. Hesselberth JR, Chen X, Zhang Z et al (2009) Global mapping of protein-DNA interactions *in vivo* by digital genomic footprinting. Nat Methods 6(4):283–289
52. Cao Y, Yao Z, Sarkar D et al (2010) Genome-wide MyoD binding in skeletal muscle cells: a potential for broad cellular reprogramming. Dev Cell 18(4):662–674
53. Fisher WW, Li JJ, Hammonds AS et al (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. Proc Natl Acad Sci U S A 109(52):21330–21335
54. Li XY, MacArthur S, Bourgon R et al (2008) Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm. PLoS Biol 6(2):e27
55. Carroll SB (2008) Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell 134(1):25–36
56. Boyle AP, Davis S, Shulha HP et al (2008) High-resolution mapping and characterization of open chromatin across the genome. Cell 132(2):311–322
57. Giresi PG, Kim J, McDaniell RM et al (2007) FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res 17(6):877–885
58. Giresi PG, Lieb JD (2009) Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). Methods 48(3):233–239
59. Buenrostro JD, Giresi PG, Zaba LC et al (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10(12):1213–1218
60. Whyte WA, Orlando DA, Hnisz D et al (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153(2):307–319
61. Visel A, Blow MJ, Li Z, Zhang T et al (2009) ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854–858
62. Heintzman ND, Stuart RK, Hon G et al (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. Nat Genet 39(3):311–318
63. Heintzman ND, Hon GC, Hawkins RD et al (2009) Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459(7243):108–112
64. Arnold CD, Gerlach D, Stelzer C et al (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339(6123):1074–1077
65. Gisselbrecht SS, Barrera LA, Porsch M et al (2013) Highly parallel assays of tissue-specific enhancers in whole *Drosophila* embryos. Nat Methods 10(8):774–780

66. Murtha M, Tokcaer-Keskin Z, Tang Z et al (2014) FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat Methods 11(5):559–565
67. Dickel DE, Zhu Y, Nord AS et al (2014) Function-based identification of mammalian enhancers using site-specific integration. Nat Methods 11(5):566–571
68. Thomas S, Li XY, Sabo PJ et al (2011) Dynamic reprogramming of chromatin accessibility during *Drosophila* embryo development. Genome Biol 12(5):R43
69. Adli M, Bernstein BE (2011) Whole-genome chromatin profiling from limited numbers of cells using nano-ChIP-seq. Nat Protoc 6(10):1656–1668
70. Nagano T, Lubling Y, Stevens TJ et al (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469):59–64
71. Goh Y, Fullwood MJ, Poh HM et al (2012) Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. J Vis Exp 62:e3770
72. He B, Chen C, Teng L et al (2014) Global view of enhancer-promoter interactome in human cells. Proc Natl Acad Sci U S A 111(21):E2191–E2199
73. Wittkopp PJ (2006) Evolution of cis-regulatory sequence and function in Diptera. Heredity 97(3):139–147
74. Li L, Zhu Q, He X, Sinha S, Halfon MS (2007) Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol 8(6):R101
75. Su J, Teichmann SA, Down TA (2010) Assessing computational methods of cis-regulatory module prediction. PLoS Comput Biol 6(12):e1001020
76. Swanson CI, Schwimmer DB, Barolo S (2011) Rapid evolutionary rewiring of a structurally constrained eye enhancer. Curr Biol 21(14):1186–1196
77. Junion G, Spivakov M, Girardot C, Braun M, Gustafson EH, Birney E, Furlong EE (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell 148(3):473–486
78. Berman BP, Pfeiffer BD, Laverty TR et al (2004) Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*. Genome Biol 5(9):R61
79. Kim J, Sinha S (2010) Towards realistic benchmarks for multiple alignments of non-coding sequences. BMC Bioinf 11:54
80. Kheradpour P, Stark A, Roy S et al (2007) Reliable prediction of regulator targets using 12 Drosophila genomes. Genome Res 17(12):1919–1931
81. Drosophila 12 Genomes Consortium (2007) Evolution of genes and genomes on the Drosophila phylogeny. Nature 450(7167):203–218
82. Sieglaff DH, Dunn WA, Xie XS et al (2009) Comparative genomics allows the discovery of *cis*-regulatory elements in mosquitoes. Proc Natl Acad Sci U S A 106(9):3053–3058
83. Kim J, Cunningham R, James B et al (2010) Functional characterization of transcription factor motifs using cross-species comparison across large evolutionary distances. PLoS Comput Biol 6(1):e1000652
84. Stark A, Lin MF, Kheradpour P et al (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450(7167):219–232
85. Brody T, Rasband W, Baler K et al (2007) cis-Decoder discovers constellations of conserved DNA sequences shared among tissue-specific enhancers. Genome Biol 8(5):R75
86. Papatsenko D, Levine M (2005) Computational identification of regulatory DNAs underlying animal development. Nat Methods 2(7):529–534
87. Papaceit M, Orengo D, Juan E (2004) Sequences upstream of the homologous cis-elements of the Adh adult enhancer of *Drosophila* are required for maximal levels of *Adh* gene transcription in adults of *Scaptodrosophila lebanonensis*. Genetics 167(1):289–299
88. Gibert JM, Simpson P (2003) Evolution of cis-regulation of the proneural genes. Int J Dev Biol 47(7–8):643–651

89. Mitsialis SA, Kafatos FC (1985) Regulatory elements controlling chorion gene expression are conserved between flies and moths. Nature 317(6036):453–456
90. Langeland JA, Carroll SB (1993) Conservation of regulatory elements controlling hairy pair-rule stripe formation. Development 117(2):585–596
91. Lukowitz W, Schroder C, Glaser G et al (1994) Regulatory and coding regions of the segmentation gene hunchback are functionally conserved between *Drosophila virilis* and *Drosophila melanogaster*. Mech Dev 45(2):105–115
92. Ludwig MZ, Patel NH, Kreitman M (1998) Functional analysis of eve stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. Development 125(5):949–958
93. Wittkopp PJ, Vaccaro K, Carroll SB (2002) Evolution of yellow gene regulation and pigmentation in Drosophila. Curr Biol 12(18):1547–1556
94. Paris M, Kaplan T, Li XY et al (2013) Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. PLoS Genet 9(9):e1003748
95. Moses AM, Pollard DA, Nix DA et al (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol 2(10):e130
96. Ludwig MZ, Bergman C, Patel NH et al (2000) Evidence for stabilizing selection in a eukaryotic enhancer element. Nature 403(6769):564–567
97. Swanson CI, Evans NC, Barolo S (2010) Structural rules and complex regulatory circuitry constrain expression of a Notch- and EGFR-regulated eye enhancer. Dev Cell 18(3):359–370
98. Erceg J, Saunders TE, Girardot C et al (2014) Subtle changes in motif positioning cause tissue-specific effects on robustness of an enhancer's activity. PLoS Genet 10(1):e1004060
99. Richards S, Liu Y, Bettencourt BR et al (2005) Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15(1):1–18
100. Hare EE, Peterson BK, Iyer VN et al (2008) Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet 4(6):e1000106
101. Cande J, Goltsev Y, Levine MS (2009) Conservation of enhancer location in divergent insects. Proc Natl Acad Sci U S A 106(34):14414–14419
102. Ludwig MZ, Palsson A, Alekseeva E et al (2005) Functional evolution of a *cis*-regulatory module. PLoS Biol 3(4):e93
103. Davidson EH (2006) The regulatory genome: gene regulatory networks in development and evolution. Academic, Burlington/San Diego
104. Stormo GD (2000) DNA binding sites: representation and discovery. Bioinformatics 16(1):16–23
105. Marinescu VD, Kohane IS, Riva A (2005) MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. BMC Bioinf 6:79
106. Cave JW, Loh F, Surpris JW et al (2005) A DNA transcription code for cell-specific gene activation by notch signaling. Curr Biol 15(2):94–104
107. Wasserman WW, Fickett JW (1998) Identification of regulatory regions which confer muscle-specific gene expression. J Mol Biol 278(1):167–181
108. Rebeiz M, Reeves NL, Posakony JW (2002) SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. Proc Natl Acad Sci U S A 99(15):9888–9893
109. Berman BP, Nibu Y, Pfeiffer BD et al (2002) Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the *Drosophila* genome. Proc Natl Acad Sci U S A 99(2):757–762
110. Rajewsky N, Vergassola M, Gaul U et al (2002) Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo. BMC Bioinf 3:30
111. Markstein M, Markstein P, Markstein V et al (2002) Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo. Proc Natl Acad Sci U S A 99(2):763–768

112. Halfon MS, Grad Y, Church GM et al (2002) Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model. Genome Res 12(7):1019–1028
113. Schroeder MD, Pearce M, Fak J et al (2004) Transcriptional control in the segmentation gene network of *Drosophila*. PLoS Biol 2(9):E271
114. Grad YH, Roth FP, Halfon MS et al (2004) Prediction of similarly acting cis-regulatory modules by subsequence profiling and comparative genomics in *Drosophila melanogaster* and *D. pseudoobscura*. Bioinformatics 20(16):2738–2750
115. Sinha S, Schroeder MD, Unnerstall U et al (2004) Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in *Drosophila*. BMC Bioinf 5:129
116. Ivan A, Halfon MS, Sinha S (2008) Computational discovery of cis-regulatory modules in *Drosophila* without prior knowledge of motifs. Genome Biol 9(1):R22
117. Kantorovitz MR, Kazemian M, Kinston S et al (2009) Motif-blind, genome-wide discovery of cis-regulatory modules in *Drosophila* and mouse. Dev Cell 17(4):568–579
118. Tompa M, Li N, Bailey TL et al (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nat Biotechnol 23(1):137–144
119. Halfon MS, Zhu Q, Brennan ER et al (2011) Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. BMC Genomics 12:578
120. Kahana S, Pnueli L, Kainth P et al (2010) Functional dissection of IME1 transcription using quantitative promoter-reporter screening. Genetics 186(3):829–841
121. Sinha S, He X (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol 3(11):e216
122. He X, Ling X, Sinha S (2009) Alignment and prediction of cis-regulatory modules based on a probabilistic model of evolution. PLoS Comput Biol 5(3):e1000299
123. Majoros WH, Ohler U (2010) Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. PLoS Comput Biol 6(12):e1001037
124. Zhou Q, Wong WH (2004) CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling. Proc Natl Acad Sci U S A 101(33):12114–12119
125. Arunachalam M, Jayasurya K, Tomancak P et al (2010) An alignment-free method to identify candidate orthologous enhancers in multiple *Drosophila* genomes. Bioinformatics 26(17):2109–2115
126. Kazemian M, Zhu Q, Halfon MS et al (2011) Improved accuracy of supervised CRM discovery with interpolated Markov models and cross-species comparison. Nucleic Acids Res 39(22):9463–9472
127. Wolff C, Schroder R, Schulz C et al (1998) Regulation of the *Tribolium* homologues of caudal and hunchback in *Drosophila*: evidence for maternal gradient systems in a short germ embryo. Development 125(18):3645–3654
128. Erives A, Levine M (2004) Coordinate enhancers share common organizational features in the Drosophila genome. Proc Natl Acad Sci U S A 101(11):3851–3856
129. Zinzen RP, Cande J, Ronshaugen M et al (2006) Evolution of the ventral midline in insect embryos. Dev Cell 11(6):895–902
130. Goltsev Y, Fuse N, Frasch M et al (2007) Evolution of the dorsal-ventral patterning network in the mosquito, *Anopheles gambiae*. Development 134(13):2415–2424
131. Kazemian M, Suryamohan K, Chen JY et al (2014) Evidence for deep regulatory similarities in early developmental programs across highly diverged insects. Genome Biol Evol 6(9):2301–2320
132. Zdobnov EM, Bork P (2007) Quantification of insect genome divergence. Trends Genet 23(1):16–20
133. FlyBase Consortium (2002) The FlyBase database of the *Drosophila* genome projects and community literature. Nucleic Acids Res 30(1):106–108

134. Munoz-Torres MC, Reese JT, Childers CP et al (2011) Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera. Nucleic Acids Res 39(Database issue):D658–D662
135. Legeai F, Shigenobu S, Gauthier JP et al (2010) AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome. Insect Mol Biol 19(Suppl 2):5–12
136. Wang L, Wang S, Li Y et al (2007) BeetleBase: the model organism database for *Tribolium castaneum*. Nucleic Acids Res 35(Database issue):D476–D479
137. Papanicolaou A, Gebauer-Jung S, Blaxter ML et al (2008) ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Res 36(Database issue):D582–D587
138. Wang J, Xia Q, He X et al (2005) SilkDB: a knowledgebase for silkworm biology and genomics. Nucleic Acids Res 33(Database issue):D399–D402
139. Negre V, Hotelier T, Volkoff AN et al (2006) SPODOBASE: an EST database for the lepidopteran crop pest *Spodoptera*. BMC Bioinf 7:322
140. Megy K, Emrich SJ, Lawson D et al (2012) VectorBase: improvements to a bioinformatics resource for invertebrate vector genomics. Nucleic Acids Res 40(Database issue): D729–D734
141. Ma Z, Yu J, Kang L (2006) LocustDB: a relational database for the transcriptome and biology of the migratory locust (*Locusta migratoria*). BMC Genomics 7:11
142. Papanicolaou A, Heckel DG (2010) The GMOD Drupal bioinformatic server framework. Bioinformatics 26(24):3119–3124
143. Lawson D, Arensburger P, Atkinson P et al (2009) VectorBase: a data resource for invertebrate vector genomics. Nucleic Acids Res 37(Database issue):D583–D587
144. Kasprzyk A (2011) BioMart: driving a paradigm change in biological data management. Database 2011:bar049
145. Karolchik D, Hinrichs AS, Kent WJ (2009) The UCSC Genome Browser. Curr Protoc Bioinformatics 40:1.4:1.r.1-4.33
146. Lyne R, Smith R, Rutherford K, Wakeling M et al (2007) FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics. Genome Biol 8(7):R129
147. Hammonds AS, Bristow CA, Fisher WW et al (2013) Spatial expression of transcription factors in *Drosophila* embryonic organ development. Genome Biol 14(12):R140
148. Tomancak P, Beaton A, Weiszmann R et al (2002) Systematic determination of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 3(12):RESEARCH0088
149. Tomancak P, Berman BP, Beaton A et al (2007) Global analysis of patterns of gene expression during *Drosophila* embryogenesis. Genome Biol 8(7):R145
150. Kumar S, Konikoff C, Van Emden B et al (2011) FlyExpress: visual mining of spatiotemporal patterns for genes and publications in *Drosophila* embryogenesis. Bioinformatics 27(23):3319–3320
151. Robinson SW, Herzyk P, Dow JA et al (2013) FlyAtlas: database of gene expression in the tissues of *Drosophila melanogaster*. Nucleic Acids Res 41(Database issue):D744–D750
152. modENCODE Consortium (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330(6012):1787–1797
153. Gallo SM, Gerrard DT, Miner D et al (2011) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. Nucleic Acids Res 39(Database issue):D118–D123
154. Sandelin A, Alkema W, Engstrom P et al (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res 32(Database issue):D91–D94
155. Wingender E, Chen X, Hehl R et al (2000) TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 28(1):316–319
156. Zhu LJ, Christensen RG, Kazemian M et al (2011) FlyFactorSurvey: a database of *Drosophila* transcription factor binding specificities determined using the bacterial one-hybrid system. Nucleic Acids Res 39(Database issue):D111–D117
157. Kazemian M, Brodsky MH, Sinha S (2011) Genome Surveyor 2.0: cis-regulatory analysis in Drosophila. Nucleic Acids Res 39(Web Server issue):W79–W85

158. Adryan B, Teichmann SA (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. Bioinformatics 22(12):1532–1533
159. Shazman S, Lee H, Socol Y et al (2014) OnTheFly: a database of *Drosophila melanogaster* transcription factors and their binding sites. Nucleic Acids Res 42(Database issue): D167–D171
160. Giardine B, Riemer C, Hardison RC et al (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455
161. Blankenberg D, Von Kuster G, Coraor N et al (2010) Galaxy: a web-based genome analysis tool for experimentalists. Curr Protoc Mol Biol 89:19.10:19.10.1–19.10.21
162. Goecks J, Nekrutenko A, Taylor J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86
163. Blankenberg D, Coraor N, Von Kuster G et al (2011) Integrating diverse databases into an unified analysis framework: a Galaxy approach. Database 2011:bar011
164. Caspari EW, Nawa S (1965) A method to demonstrate transformation in *Ephestia*. Z Naturforsch 20b:281–284
165. Germeraad S (1976) Genetic transformation in *Drosophila* by microinjection of DNA. Nature 262(5565):229–231
166. Rubin GM, Spradling AC (1982) Genetic transformation of *Drosophila* with transposable element vectors. Science 218(4570):348–353
167. Spradling AC, Rubin GM (1982) Transposition of cloned P elements into *Drosophila* germ line chromosomes. Science 218(4570):341–347
168. Loukeris TG, Livadaras I, Arca B et al (1995) Gene transfer into the medfly, *Ceratitis capitata*, with a *Drosophila hydei* transposable element. Science 270(5244):2002–2005
169. Pavlopoulos A, Oehler S, Kapetanaki MG et al (2007) The DNA transposon Minos as a tool for transgenesis and functional genomic analysis in vertebrates and invertebrates. Genome Biol 8(Suppl 1):S2
170. Handler AM, McCombs SD, Fraser MJ et al (1998) The lepidopteran transposon vector, piggyBac, mediates germ-line transformation in the Mediterranean fruit fly. Proc Natl Acad Sci U S A 95(13):7520–7525
171. Balu B, Shoue DA, Fraser MJ Jr et al (2005) High-efficiency transformation of *Plasmodium falciparum* by the lepidopteran transposable element piggyBac. Proc Natl Acad Sci U S A 102(45):16391–16396
172. Handler AM (2002) Use of the piggyBac transposon for germ-line transformation of insects. Insect Biochem Mol Biol 32(10):1211–1220
173. Wilson MH, Coates CJ, George AL Jr (2007) PiggyBac transposon-mediated gene transfer in human cells. Mol Ther 15(1):139–145
174. Berghammer A, Bucher G, Maderspacher F et al (1999) A system to efficiently maintain embryonic lethal mutations in the flour beetle *Tribolium castaneum*. Dev Genes Evol 209(6):382–389
175. Schinko JB, Weber M, Viktorinova I et al (2010) Functionality of the GAL4/UAS system in *Tribolium* requires the use of endogenous core promoters. BMC Dev Biol 10:53
176. Berghammer AJ, Weber M, Trauner J et al (2009) Red flour beetle (*Tribolium*) germline transformation and insertional mutagenesis. Cold Spring Harb Protoc 2009(8):pdb prot5259
177. Trauner J, Schinko J, Lorenzen MD et al (2009) Large-scale insertional mutagenesis of a coleopteran stored grain pest, the red flour beetle *Tribolium castaneum*, identifies embryonic lethal mutations and enhancer traps. BMC Biol 7:73
178. Eckert C, Aranda M, Wolff C et al (2004) Separable stripe enhancer elements for the pair-rule gene hairy in the beetle *Tribolium*. EMBO Rep 5(6):638–642
179. Uchino K, Sezutsu H, Imamura M et al (2008) Construction of a piggyBac-based enhancer trap system for the analysis of gene function in silkworm *Bombyx mori*. Insect Biochem Mol Biol 38(12):1165–1173
180. Marcus JM, Ramos DM, Monteiro A (2004) Germline transformation of the butterfly *Bicyclus anynana*. Proc Roy Soc 271(Suppl 5):S263–S265

181. Schulte C, Theilenberg E, Muller-Borg M et al (2014) Highly efficient integration and expression of piggyBac-derived cassettes in the honeybee (*Apis mellifera*). Proc Natl Acad Sci U S A 111(24):9003–9008
182. Groth AC, Fish M, Nusse R et al (2004) Construction of transgenic *Drosophila* by using the site-specific integrase from phage phiC31. Genetics 166(4):1775–1782
183. Oberstein A, Pare A, Kaplan L et al (2005) Site-specific transgenesis by Cre-mediated recombination in *Drosophila*. Nat Methods 2(8):583–585
184. Horn C, Handler AM (2005) Site-specific genomic targeting in *Drosophila*. Proc Natl Acad Sci U S A 102(35):12483–12488
185. Bateman JR, Lee AM, Wu CT (2006) Site-specific transformation of *Drosophila* via phiC31 integrase-mediated cassette exchange. Genetics 173(2):769–777
186. Venken KJ, He Y, Hoskins RA et al (2006) P[acman]: a BAC transgenic platform for targeted insertion of large DNA fragments in *D. melanogaster*. Science 314(5806):1747–1751
187. Bischof J, Maeda RK, Hediger M et al (2007) An optimized transgenesis system for *Drosophila* using germ-line-specific phiC31 integrases. Proc Natl Acad Sci U S A 104(9):3312–3317
188. Amenya DA, Bonizzoni M, Isaacs AT et al (2010) Comparative fitness assessment of *Anopheles stephensi* transgenic lines receptive to site-specific integration. Insect Mol Biol 19(2):263–269
189. Meredith JM, Basu S, Nimmo DD et al (2011) Site-specific integration and expression of an anti-malarial gene in transgenic *Anopheles gambiae* significantly reduces *Plasmodium* infections. PLoS One 6(1):e14587
190. Nakayama G, Kawaguchi Y, Koga K et al (2006) Site-specific gene integration in cultured silkworm cells mediated by phiC31 integrase. Mol Genet Genomics 275(1):1–8
191. Nimmo DD, Alphey L, Meredith JM et al (2006) High efficiency site-specific genetic engineering of the mosquito genome. Insect Mol Biol 15(2):129–136
192. Schetelig MF, Scolari F, Handler AM et al (2009) Site-specific recombination for the modification of transgenic strains of the Mediterranean fruit fly *Ceratitis capitata*. Proc Natl Acad Sci U S A 106(43):18171–18176
193. Labbe GM, Nimmo DD, Alphey L (2010) piggybac- and PhiC31-mediated genetic transformation of the Asian tiger mosquito, *Aedes albopictus* (Skuse). PLoS Negl Trop Dis 4(8):e788
194. Fu G, Lees RS, Nimmo D et al (2010) Female-specific flightless phenotype for mosquito control. Proc Natl Acad Sci U S A 107(10):4550–4554
195. O'Brochta DA, Pilitt KL, Harrell RA 2nd et al (2012) Gal4-based enhancer-trapping in the malaria mosquito *Anopheles stephensi*. G3 2(11):1305–1315
196. Adelman ZN, Jasinskiene N, Onal S et al (2007) nanos gene control DNA mediates developmentally regulated transposition in the yellow fever mosquito *Aedes aegypti*. Proc Natl Acad Sci U S A 104(24):9970–9975
197. Wise de Valdez MR, Nimmo D, Betz J et al (2011) Genetic elimination of dengue vector mosquitoes. Proc Natl Acad Sci U S A 108(12):4772–4775
198. Griffiths JR, Salanitri VR (1980) The strength of spider silk. J Mater Sci 15(2):491–496
199. Wen H, Lan X, Zhang Y et al (2010) Transgenic silkworms (*Bombyx mori)* produce recombinant spider dragline silk in cocoons. Mol Biol Rep 37(4):1815–1821
200. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8(3):206–216
201. Carroll SB (2005) Evolution at two levels: on genes and form. PLoS Biol 3(7):e245
202. Whitehead A, Crawford DL (2006) Variation within and among species in gene expression: raw material for evolution. Mol Ecol 15(5):1197–1211
203. Rubinstein M, de Souza FS (2013) Evolution of transcriptional enhancers and animal diversity. Philos Trans R Soc Lond B Biol Sci 368(1632):20130017
204. Sucena E, Stern DL (2000) Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by cis-regulatory evolution of *ovo/shaven-baby*. Proc Natl Acad Sci U S A 97(9):4530–4534

205. Carbone MA, Llopart A, deAngelis M et al (2005) Quantitative trait loci affecting the difference in pigmentation between *Drosophila yakuba* and *D. santomea*. Genetics 171(1):211–225
206. Pool JE, Aquadro CF (2007) The genetic basis of adaptive pigmentation variation in *Drosophila melanogaster*. Mol Ecol 16(14):2844–2851
207. Frankel N, Erezyilmaz DF, McGregor AP et al (2011) Morphological evolution caused by many subtle-effect substitutions in regulatory DNA. Nature 474(7353):598–603
208. Stone JR, Wray GA (2001) Rapid evolution of cis-regulatory sequences via local point mutations. Mol Biol Evol 18(9):1764–1770
209. Deng Q, Ramskold D, Reinius B et al (2014) Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. Science 343(6167):193–196
210. Pennacchio LA, Bickmore W, Dean A et al (2013) Enhancers: five essential questions. Nat Rev Genet 14(4):288–295
211. D'Haeseleer P (2006) What are DNA sequence motifs? Nat Biotechnol 24(4):423–425
212. Zambelli F, Pesole G, Pavesi G (2013) Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. Brief Bioinform 14(2):225–237
213. MacIsaac KD, Fraenkel E (2006) Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol 2(4):e36
214. Haeussler M, Joly JS (2011) When needles look like hay: how to find tissue-specific enhancers in model organism genomes. Dev Biol 350(2):239–254
215. Van Loo P, Marynen P (2009) Computational methods for the detection of cis-regulatory modules. Brief Bioinform 10(5):509–524
216. Aerts S (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr Top Dev Biol 98:121–145
217. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nat Methods 5(1):19–21
218. Zentner GE, Scacheri PC (2012) The chromatin fingerprint of gene enhancer elements. J Biol Chem 287(37):30888–30896
219. Shlyueva D, Stampfel G, Stark A (2014) Transcriptional enhancers: from properties to genome-wide predictions. Nat Rev Genet 15(4):272–286
220. Mohr SE, Hu Y, Kim K et al (2014) Resources for functional genomics studies in *Drosophila melanogaster*. Genetics 197(1):1–18
221. Prud'homme B, Gompel N, Carroll SB (2007) Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A 104(Suppl 1):8605–8612
222. Carroll SB, Grenier JK, Weatherbee SD (2005) From DNA to diversity: molecular genetics and the evolution of animal design, 2nd edn. Blackwell, Malden

Chapter 7
Comparative Genomics of Transcription Factor Binding in *Drosophila*

Sarah Carl and Steven Russell

Abstract While the number of genome-wide, in vivo transcription factor binding datasets is growing, yielding greater insight into the role of regulatory DNA in development, evolution and disease, it is difficult to tease apart signal from noise and identify truly functional binding events. Comparative studies of transcription factor binding between closely related species offer one way to combat this problem, as functionally important aspects of enhancer architecture tend to be constrained by natural selection. Here we review the current field in the area of in vivo transcription factor binding in *Drosophila*, illustrating how evolutionary studies within the drosophilids are helping to unravel the complexity of the genomic regulatory code. A number of techniques exist for studying transcription factor binding on a genome-wide scale, including ChIP-chip, ChIP-seq and DamID; we touch on these and address the challenges and advantages of each with regard to working on non-model species. We also describe major findings in the field so far, focusing on comparative studies of the developmental regulatory network, the logic of combinatorial binding and the evolutionary properties of noncoding DNA. Finally, we examine how insights from *Drosophila* compare with similar studies in the vertebrates and address some open questions that have been raised by studies conducted thus far.

Abbreviations

A-P	Anterior-posterior
ChIP	Chromatin immunoprecipitation
ChIP-chip	Chromatin immunoprecipitation combined with the use of microarray chips
ChIP-seq	Chromatin immunoprecipitation combined with parallel array sequencing
Dam	DNA adenine methyltransferase
DamID	DNA adenine methyltransferase identification

S. Carl • S. Russell (✉)
Department of Genetics and Cambridge Systems Biology Centre, University of Cambridge, Downing Street, Cambridge CB2 3EH, UK
e-mail: sarahhcarl@gmail.com; s.russell@gen.cam.ac.uk

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*, Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_7

DamID-seq	DNA adenine methyltransferase identification combined with parallel array sequencing
eQTL	Expression quantitative trait locus
FWOB	Four-way orthologous binding
GFP	Green fluorescent protein
GO	Gene ontology
modENCODE	Model Organism Encyclopedia of DNA Elements
ORF	Open reading frame
PCA	Principle component analysis
PWM	Positional weight matrix
RNA-seq	RNA sequencing
TF	Transcription factor
TWOB	Two-way orthologous binding
UCSC	University of California, Santa Cruz
UAS	Upstream activating sequence

7.1 Introduction

The importance of regulatory DNA in development, disease and evolution is widely accepted and becoming a key focus for genomics as large-scale studies such as the ENCODE project attempt to map diverse elements of the noncoding genome [1–4]. One of the major roles of regulatory DNA is to bind transcription factors and, together with other genomic elements such as promoters, to direct gene expression in a temporally and spatially specific manner. In the model organism, *Drosophila melanogaster*, significant strides have been made towards understanding how multiple inputs are integrated to determine transcription factor occupancy in the nucleus and how, in turn, combinatorial rules of transcription factor binding describe functional regulatory elements [5–7]. However, the primary methods for determining transcription factor binding, both in vivo and in silico, suffer from difficulties in distinguishing between true functional events and biological noise, resulting in high numbers of potential false positives and making it difficult to tease apart underlying regulatory networks [8–10]. Comparative studies of transcription factor binding in multiple *Drosophila* species facilitate the use of patterns of conservation to identify functional features of the regulatory genome as well as an analysis of the evolutionary dynamics of transcriptional regulation.

One of the first comparative studies of regulatory DNA in *Drosophila* was an in-depth dissection of the *even-skipped stripe 2* enhancer in six *Drosophila* species and six species of sepsid flies by Hare et al. [11]. Using both computational prediction of transcription factor binding sites and in vitro footprinting, this pioneering study made the surprising observation that highly diverged regulatory architectures could nonetheless lead to very similar phenotypic outputs. However, it also showed a high degree of conservation in transcription factor binding sites between *Drosophila* species despite significant sequence divergence overall, suggesting that

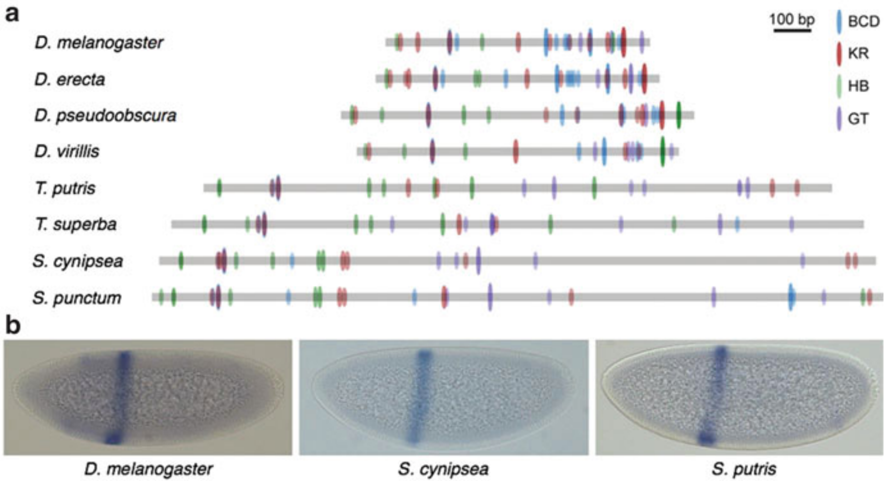


Fig. 7.1 (**a**) Predicted binding sites for the segmentation transcription factors Bicoid (BCD), Kruppel (KR), Hunchback (HB) and Giant (GT) in the *even-skipped stripe 2 enhancers* from four *Drosophila* and four sepsid species (*Themira putris*, *Themira superba*, *Sepsis cynips*ea and *Sepsis punctum*). The height and colour intensity of each oval represent the degree of similarity to a position weight matrix for each TF [66]. (**b**) The activity of the even-skipped stripe 2 enhancers from the indicated species as assayed by reporter constructs in *D. melanogaster* embryos. Even though the enhancer sequence and arrangement of TF binding sites differ between the species, they produce very similar transcriptional output [11]. Both figures reproduced under the Creative Commons Attribution License

binding patterns are important in determining the functionality of regulatory elements (Fig. 7.1) [11].

Detailed genetic studies have also examined loci involved in wing and body pigmentation, in particular the *yellow* pigmentation gene, and also trichome patterning on the larval cuticle via the *shaven baby* (*svb*) gene [12–14]. The advantages of these targeted studies include the ability to identify and dissect well-characterised orthologous enhancers at a sequence level between species and to validate predictions by testing sequences from different species, or constructs with engineered mutations, in transgenic *D. melanogaster* assays. Using such methods, it is possible to closely correlate sequence with phenotype and identify causal mutations as well as changes that occur repeatedly during evolution. Despite the utility of these focused studies, newly developed techniques that assay genome-wide patterns of transcription factor binding in vivo offer an unbiased approach to discovering and comparing regulatory elements in multiple species. The genomic datasets generated by such approaches facilitate the identification of functionally important characteristics of the behaviour of classical transcription factors and other DNA-binding proteins, such as insulators, as well as the DNA motifs they recognise and the local chromatin environments surrounding regulatory elements. In the following sections, we will outline the techniques available for measuring transcription factor binding in vivo in multiple species of *Drosophila*, highlighting advantages and challenges of each, and then

discuss the results of several comparative studies that have been performed so far, focusing on the implications for functional gene regulation.

7.2 Methods for Assaying Genome-Wide Binding

A number of different techniques exist for directly or indirectly studying genome-wide transcription factor binding patterns in *Drosophila*, each of which present different advantages and challenges for use in non-model species. Two of the primary in vivo techniques are chromatin immunoprecipitation (ChIP) and DNA adenine methyltransferase identification (DamID), which is based on DNA methylation by a tethered DNA adenine methyltransferase (Dam) [15]. Each of these techniques can be combined with either hybridization to a tiling microarray or high-throughput sequencing to identify preferentially bound regions genome wide [16, 17]. However, since microarrays are generally not commercially available for non-model species and the cost of sequencing has dropped significantly in the last decade, sequencing has become the method of choice for most comparative studies. Since ChIP-seq and DamID-seq use fundamentally different biological mechanisms to ask the same questions, they can be used to validate and complement one another [17]. However, certain key differences between the two techniques necessitate different data analysis methods and should raise caution in the interpretation of results.

The most popular technique for comparative studies of transcription factor binding in *Drosophila* at the present moment is ChIP-seq (Fig. 7.2a). With the publication of the modENCODE data in 2010 [18], a large number of ChIP-chip (chromatin immunoprecipitation combined with data from microarray “chips”) and ChIP-seq (chromatin immunoprecipitation combined with high-throughput parallel array sequencing) datasets from *Drosophila melanogaster* were made publicly available. At the time of writing, the modMine database, which houses the modENCODE datasets, contains 279 entries for ChIP-chip and ChIP-seq datasets for transcription factor binding as well as chromosomal proteins and histone modifications in *D. melanogaster* [19]. In addition, a more focused study on the binding of 31 transcription factors involved in early embryonic patterning, along with matching chromatin accessibility data, are available via the Berkeley *Drosophila* Transcription Network Project [10]. The availability of these datasets, as well as data-processing tools, quality control guidelines and experimental best practice guidance from the modENCODE consortium [20, 21], provides a valuable resource for researchers wishing to undertake comparative studies in other *Drosophila* species. ChIP-seq experiments have been successfully performed with transcription factors in *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura* and *D. virilis* [22–25], representing an evolutionary span of approximately 40 Ma. However, ChIP-seq in non-model species is subject to some limitations.

One of the key factors determining the success or failure of a ChIP-seq experiment is the quality of the antibody used; the modENCODE ChIP-seq guidelines contain extensive information on tests for the characterisation of new antibodies

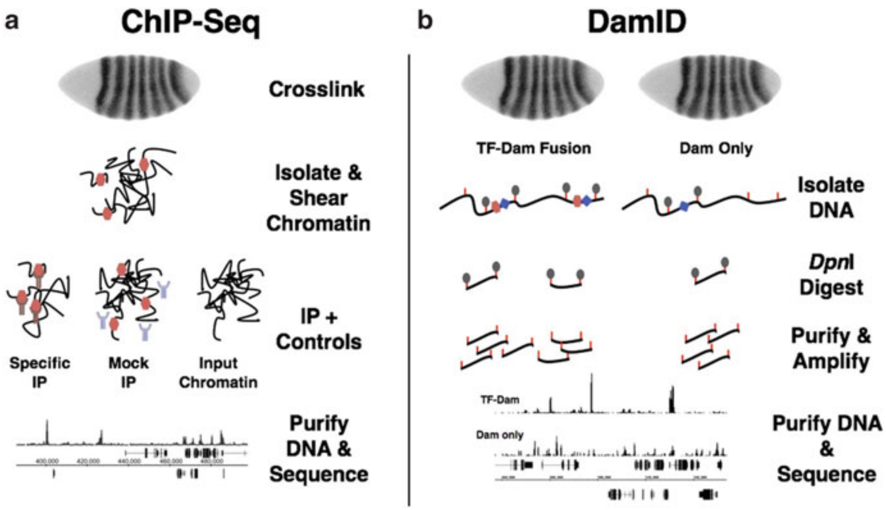


Fig. 7.2 Methods for mapping in vivo DNA binding. (**a**) A typical ChIP-seq pipeline: embryos are cross-linked with formaldehyde prior to chromatin isolation and fragmentation. The chromatin sample is enriched with an antibody specific to the TF of interest; mock IP and input chromatin controls are prepared in parallel. Processed sequence reads are mapped to the reference genome. (**b**) In the DamID approach, two transgenic lines are generated, one expressing a TF-dam fusion and a second expressing a Dam-only control. DNA isolated from each line is subject to DpnI digestion, which cuts GATC sequences when the A is methylated. After processing, sequence reads are mapped to the reference genome – the *top trace* represents the Dam-TF fusion and the lower the Dam-only control. TFs are represented by *hexagons*, specific antibodies in *dark grey* and control antibodies (i.e., IgG) in *light grey*. The Dam enzyme is represented by a *diamond*, *vertical lines* represent GATC motifs and the *ovals* represent methylated A residues

[20]. When comparing closely related species, it is often assumed that specific antibodies from one species will cross-react sufficiently with the respective orthologous proteins; however, while this can be assayed to some extent by immunohistochemistry, it does not always translate to ChIP efficiency. Affinity purification of antibodies raised against proteins from one species using proteins from related species may help increase specificity and avoid biases due to antibodies recognising more epitopes for the protein in the species against which they were raised [24]. However, identifying ChIP-quality antibodies can be difficult for any experiment and is generally more difficult in non-model species, as commercially available antibodies normally do not exist for such species. One alternative is to use antibodies recognising a specific tag, such as green fluorescent protein (GFP) [26], and to tag the transcription factor of interest in the species to be examined. However, this approach requires the creation of transgenic lines, which can be time- and labour-intensive, and especially in the case of transcription factors, it may not be possible to express the tagged protein under endogenous regulatory controls. Nonetheless, if a suitable antibody can be found, the major advantage of ChIP-seq is that it directly measures binding of endogenous proteins in a native temporal and spatial context.

DamID presents an alternative method to chromatin precipitation for identifying transcription factor binding (Fig. 7.2b). The basic principle of DamID is that a DNA adenine methyltransferase (Dam) from *E. coli* is fused to the transcription factor of interest, which is then expressed at low levels in a transgenic organism. In *Drosophila*, this can be achieved via leaky expression in constructs carrying un-induced *UAS* sites and a minimal *hsp70* promoter. The Dam enzyme methylates adenine residues at GATC sites surrounding the genomic loci bound by the fusion protein. Adenine methylation is not endogenously present in eukaryotic organisms, allowing bound fragments to be isolated by a series of digestions with the methylation-sensitive restriction enzyme DpnI and the methylation-insensitive isoschizomer DpnII. Due to Dam's high affinity for DNA, a Dam-only construct is expressed in parallel to control for the detection of nonspecific methylation [15, 27]. Although DamID is not as commonly used as ChIP, it has been successfully implemented in *Drosophila*, yeast, *Arabidopsis* and mammalian cells and has recently been used for comparative binding studies in *Drosophila* species [28].

One advantage of DamID in comparison to ChIP is that it does not rely on the availability of a high-quality antibody; as such, it can be performed for any protein whose sequence is known. DamID does require the generation of transgenic lines, which may not be as straightforward in non-model species as it is in *D. melanogaster*. Nonetheless, transgenic technology is possible in other species: in particular *PiggyBac*-based vectors have been shown to facilitate transformation in a wide range of insects, including various drosophilids [29]. The main difficulties in working with non-model species include the lack of balancer chromosomes, making it difficult to establish homozygous lines, and the lack of site-specific recombination systems, such as phiC31-mediated integration [30], meaning that transgene insertion sites are random and potentially vulnerable to positional effects. Until recently, another potential drawback of DamID was that the fusion protein could only be expressed at low levels throughout the organism; as such, temporal and spatial specificity were lost. A new technique allows for targeted expression of Dam fusion proteins ("TaDa") through the use of a bicistronic transcript encoding the Dam fusion as a secondary open reading frame (ORF) whose translation can be specifically induced using GAL4 [31]. This allows for more specific assays of transcription factor binding, for example, in a particular tissue at a particular stage of development, although it does require more complex transgenesis since it relies on a two-component system.

Another consideration when performing comparative DamID experiments is whether to use the endogenous protein in each species being studied, which requires the construction of separate vectors, or whether to use one fusion protein for all species. It is certainly possible to clone each orthologous protein separately, and indeed, this approach would capture binding events that may be closer to the native state for each species. While many families of transcription factors are highly conserved between closely related species of *Drosophila*, particularly in their DNA-binding domains [32], and are likely to recognise the same consensus sequences, small changes in amino acid sequences can alter DNA-binding properties [33]. Bearing in mind that all DamID experiments involve expressing a transgenic fusion

protein alongside the endogenous transcription factor, for some experiments it might be desirable to express a fusion protein based on the protein of interest from the reference species in each of the other species studied. This approach would subtly change the interpretation of the experiment, as any detected variation in binding would be entirely attributable to changes in *cis* (i.e., in the DNA sequence and local chromatin landscape), rather than due to changes in binding preferences of the protein itself. Therefore, such an experiment would answer a different but also interesting question in comparison to ChIP-seq experiments.

In addition to directly measuring transcription factor binding in vivo, computational methods can be used to predict binding sites in multiple species, search for signatures of selection and understand conserved features of regulatory elements. Often these techniques involve detecting binding intervals in vivo in one species, searching for bound sequence motifs in those intervals, and then using a positional weight matrix (PWM) constructed from the discovered motifs to scan other genomes for orthologous sites. While this approach cannot guarantee that orthologous sites are actually bound in vivo, it does allow for a sequence-level analysis of putative regulatory elements in multiple species. A variety of computational methods have been developed to address the challenge of identifying orthologous enhancers, which can be confounded by alignment gaps and genome rearrangements [34–36]. For example, the evolutionary analysis of noncoding DNA from the genomes of 12 sequenced *Drosophila* species led to the prediction of regulatory regions as well as motifs for individual transcription factors [37–39]. In some cases, these predictions have been validated in functional transgenic assays; however, this approach is quite time-consuming, and some evidence suggests that a large proportion of the individual binding sites predicted solely on the basis of sequence conservation may not be functional in vivo [40]. In cases where it is available, integrating in vivo binding data is seen as preferable to simply using computational predictions of binding sites, since it can capture spatial and temporal variation in the nuclear environment and provide a non-biased functional validation of predictions.

7.3 Calculating Constraint and Turnover

One of the most fundamental questions that comparative transcription factor binding studies can ask is whether, and to what extent, individual binding events are conserved between different species (Fig. 7.3). However, binding conservation is a complex phenomenon and can be examined on multiple levels. ChIP-seq and DamID-seq experiments typically produce binding profiles, consisting of counts of sequence reads that align to the genome. Analysis of the sequence data produces sets of called peaks or binding intervals, which often have associated scores based on the number of reads aligning to a given position (sometimes referred to as the intensity or height of a peak) or the probability of a position being enriched in a bound sample in comparison to a control sample or a background model. Functional genomic studies are often focused on compiling a list of target genes for a particular

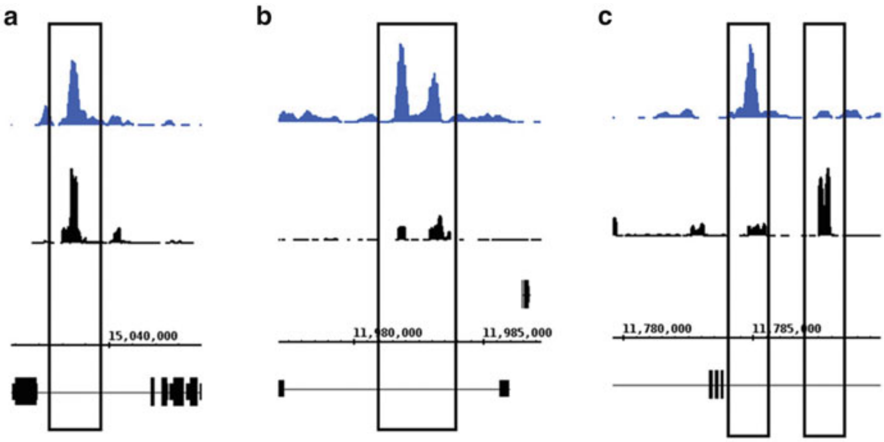


Fig. 7.3 Examples of conserved and differential binding profiles. (**a**) A peak conserved in both position and intensity. (**b**) A pair of peaks that have decreased intensity at the same position in one species. (**c**) Peaks of similar intensity shifted in one species but in close proximity to the reference species. *Top traces* represent Dichaete-DamID from *D. melanogaster*; *bottom traces* represent Dichaete-DamID in *D. simulans* [28]

factor, and in such cases the binding peaks are assigned to the genes they putatively regulate. Assigning binding peaks to regulated genes can be complicated by the fact that *cis*-regulatory elements may be located upstream or downstream from the genes they regulate or reside within introns and do not necessarily act on the closest gene [41]. While databases of experimentally verified *cis*-regulatory elements such as *REDfly* and *ORegAnno* [42, 43] may be helpful in this regard, and the recent development of high-throughput enhancer detection assays such as STARR-seq [44] are beginning to map putative general and tissue-specific enhancers, there are currently no validated enhancer databases covering the entire *D. melanogaster* genome. Additionally, the genomes of non-model *Drosophila* species are generally not as well annotated as that of *D. melanogaster*. Nevertheless, a comparison of potential target genes can afford a high-level, functional comparison of interspecies binding.

In order to analyse conserved binding events more directly, datasets from multiple species must be remapped to the genome of a common reference species; in *Drosophila*, this is typically the *D. melanogaster* genome. Remapping can be performed either at the level of sequence reads themselves or with called peak locations [45]. One way to achieve this is using the LiftOver utility from the UCSC Genome Browser, which can translate genomic positions between different assemblies of the same genome or between closely related genomes using a global alignment algorithm [46]. From this point, a number of different comparisons can be made. At the simplest level, the peak locations from different species can be compared, typically using the BEDTools set of utilities for identifying the intersection or union of sets of interval data and calculating the Jaccard index, a measure of the degree of overlap between two entire datasets [47]. Statistical models can also be used to evaluate the

significance of overlaps between two sets of genomic intervals [48, 49], adding a quantitative dimension to the analysis. However, to obtain a more precise and detailed view of differential and conserved binding, peak scores, read counts or read densities from binding profiles should be taken into account. In all cases, the use of controls such as input chromatin and appropriate data normalisation methods is important to prevent the erroneous identification of differential peaks that are in fact due to differences in local chromatin environments or sequence composition (e.g., GC content) in the species being compared [22, 49].

7.4 Estimates of Binding Divergence and Conservation in *Drosophila*

Several studies, focusing on different transcription factors and using different sets of species, have independently attempted to estimate binding conservation as well as the rate of binding site turnover in *Drosophila*. One of the first of these used ChIP-chip to measure genome-wide binding of the transcription factor Zeste. ChIP-chip was performed only in *D. melanogaster*, and the resulting binding intervals were aligned against the genomes of D. *simulans*, *D. erecta* and *D. yakuba* [50]. Since in vivo binding data was only available for one species, an analysis of quantitative differences in binding between species was not possible; instead, the authors considered binding as a binary state based on called peaks. Using a conservative approach, only binding intervals identified in *D. melanogaster* that could be unambiguously aligned to orthologous sequences in each of the other species were included, and the analysis was further restricted to those intervals containing matches to a Zeste motif PWM. Nonetheless, the authors found that at least 5 % of Zeste binding sites identified in *D. melanogaster* were not conserved in the other species they examined, implying that those sites were either gained in the *D. melanogaster* lineage or lost in the other lineages since the divergence of the *melanogaster* subgroup [50].

Several more recent studies employing ChIP-seq or DamID-seq to measure transcription factor binding in multiple species of *Drosophila* generated broadly similar estimates of binding site conservation. One of these examined binding of 6 transcription factors involved in anterior-posterior (A-P) patterning in the early embryo, Bicoid (Bcd), Hunchback (Hb), Kruppel (Kr), Giant (Gt), Knirps (Kni) and Caudal (Cad), in the closely related species *D. melanogaster* and *D. yakuba* [22]. A subsequent experiment by the same group expanded the phylogenetic distance by measuring the binding of four of these factors (Bcd, Gt, Hb and Kr) in the same two species along with *D. pseudoobscura* and *D. virilis* [24]. A third study focused on the mesodermal regulator Twist (Twi) in six species: *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. erecta*, *D. ananassae* and *D. pseudoobscura*, which span approximately 25 Ma of evolutionary time [23]. Finally, a recent DamID-seq experiment determined the binding of the Sox family transcription factors Dichaete and SoxN in *D.*

melanogaster, *D. simulans*, *D. yakuba* and *D. pseudoobscura* [28]. Each of these studies considered both presence/absence of peaks in each species as well as quantitative changes in binding strength.

Bradley et al. found that, for each of the six factors studied, between 1 and 15 % of peaks that were identified in one species were absent in the other. They measured quantitative binding divergence by calculating the genome-wide correlations between binding strength at all peaks for each factor in *D. melanogaster* and *D. yakuba*; these values ranged from 0.57 to 0.75 for peaks at genes not known to be regulated by the anterior-posterior (A-P) patterning factors [22]. In similar pairwise comparisons between binding strengths of peaks in *D. melanogaster* and *D. pseudoobscura*, the correlations ranged from 0.37 for Gt to 0.64 for Kr, reflecting the greater phylogenetic distance between the two species [24]. An equivalent analysis performed on Dichaete and SoxN peaks identified in our recent DamID-seq study revealed correlations in binding strength between *D. melanogaster*, *D. simulans* and *D. yakuba* ranging from 0.62 to 0.72 [28], indicating an overall similarity in estimates of binding conservation for different transcription factors measured using different technologies. In the case of Twi, around 80 % of peaks identified in *D. melanogaster* were found to be conserved in *D. simulans* and *D. yakuba*, with the percentage decreasing to around 60 % for *D. pseudoobscura*. The authors measured quantitative divergence by computing the number of peaks whose binding strength changed between *D. melanogaster* and each other species; this ranged from around 10–35 % of total peaks and increased linearly with evolutionary time [23].

Finally, two studies have examined genome-wide binding patterns of the insulator proteins CTCF (in *D. melanogaster*, *D. simulans*, *D. yakuba* and *D. pseudoobscura*) and BEAF-32 (in *D. melanogaster*, *D. simulans*, *D. pseudoobscura* and *D. virilis*) [51, 52]. For CTCF, the authors considered both binding events where orthologous sequences could be identified in pairwise comparisons between species (two-way orthologous binding or TWOBs) as well as where orthologous sequences could be identified in all four species studied (four-way orthologous binding or FWOBs). They found that, of all CTCF-binding intervals identified in *D. melanogaster*, around 20 % were not present in *D. simulans* and up to 74 % were not present in *D. pseudoobscura*; the percentages of divergent sites were slightly lower for FWOBs than for TWOBs, perhaps reflecting the higher overall sequence conservation at FWOB sites [51]. The divergence rate between *D. melanogaster* and *D. simulans* is very close to that measured for Twi; however, CTCF divergence appears to increase more quickly with evolutionary time, leading to a much higher percentage of divergent sites in *D. pseudoobscura*. This is in contrast to the situation in vertebrates, in which CTCF binding appears to be more conserved than that of other transcription factors; it has been suggested that this discrepancy may be due to different evolutionary mechanisms operating in the two clades [25, 53, 54]. Somewhat surprisingly, BEAF-32 binding intervals appear to be much more conserved, with percentages of divergent sites ranging from 3 % between *D. melanogaster* and *D. simulans* to 29 % between *D. melanogaster* and *D. virilis* [52]. Although these data appear discordant with the observation that CTCF and BEAF-32 binding are highly correlated in the *Drosophila* embryo [55], caution should be applied when directly

comparing the conclusions due to the different analytical approaches used to identify conserved or divergent binding. Nevertheless, it is interesting to note that multiple insulator-binding proteins have been characterised in flies, in contrast to the situation in vertebrates, where CTCF is the only known insulator protein [51], and it is therefore possible that different patterns of CTCF- and BEAF-32-binding conservation may reflect differing levels of evolutionary constraint on the two proteins and/or specific differences in their roles and genomic localization [25].

7.5 Conservation and Function

Besides simply estimating rates of binding conservation and divergence, comparative studies of transcription factor binding can identify functional features of transcription factors by considering differences in binding conservation relative to genomic annotations or patterns of binding by other factors. This type of analysis builds on the hypothesis that functional sites will be subject to purifying selection and thus will be preferentially conserved. One way to test this hypothesis is to evaluate conservation at a set of well-characterised functional regulatory elements. For example, as we note above, peaks for A-P patterning regulators are more conserved at known A-P target genes compared to all genes, and peaks for Twi binding are highly conserved at regulatory elements that are known Twi targets [22–24]. Additionally, the most highly conserved Twi peaks show an enrichment near genes that are down-regulated in *Twi* mutants as well as genes that are annotated with gene ontology (GO) functions related to the developmental role of Twi [23], both of which are also indicators of function.

Since transcriptional regulators direct expression of their target genes, it is interesting to ask whether divergence in binding is correlated with divergence in expression levels between species. In the study of Gt, Hb, Kr and Bcd binding, this question was addressed by performing RNA-seq in each species to measure species-specific changes in gene expression [24]. It was found that mRNA levels were highly correlated between species and showed significantly less overall divergence than the binding patterns of any of the transcription factors studied. Perhaps surprisingly, although the authors found a significant positive correlation between transcription factor binding and mRNA levels for nearby genes within each species, the relationship between binding variation and gene expression between species was weak. According to this analysis, only around 2 % of the variation in mRNA levels between species could be attributed to variation in transcription factor (TF) binding at zygotically active genes [24]. However, the authors note that ChIP-seq is a much noisier technique than RNA-seq and that some of their annotations of binding intervals to target genes may have been incorrect; both of these factors could obscure the relationship between binding and gene expression to some extent. Indeed, the authors of the comparative CTCF study performed a similar RNA-seq experiment and uncovered a stronger correlation between divergence of CTCF sites and divergent mRNA levels of nearby genes [51]. It is possible that levels of CTCF binding might

have a more detectable effect on gene expression, since it is believed to be responsible for the establishment of regulatory domains across the genome, as opposed to the A-P transcription factors, which are likely to directly regulate genes at only a subset of their binding sites and which operate combinatorially with other transcription factors.

It is also possible to examine the effect of sequence-level conservation on transcription factor binding. Both the two A-P factor studies, the Twi study and our Sox study described above show that, while overall sequence conservation in bound regions does not correlate strongly with binding divergence, measured either quantitatively or as a binary presence/absence call, conservation of short sequence motifs within binding intervals does show some correlation with binding divergence [22–24, 28]. He et al. found that Twi peaks present in all four species studied had significantly more fully conserved Twi motifs than peaks that were only present in *D. melanogaster*; Carl and Russell came to the same conclusion for Dichaete motifs [23, 28]. Moreover, the quality of Twi motifs present in peaks was also correlated with quantitative changes in binding strength between species [23]. A similar conclusion was reached for BEAF-32, as peaks that were conserved in pairwise comparisons between *D. melanogaster* and *D. pseudoobscura* or *D. melanogaster* and *D. virilis* were more likely to contain conserved BEAF-32 motifs than peaks that were not present in the non-melanogaster species [52]. Interestingly, the CTCF study was the only one to find that sequence conservation over broader (200 bp) regions surrounding binding sites was significantly higher for peaks that were conserved in multiple species than for peaks unique to *D. melanogaster*, possibly indicating a more important role for genomic sequence context in CTCF binding than for other more canonical transcription factors [51].

Despite the fact that motif conservation is correlated with binding conservation, it does not explain all of the observed binding divergence in any of the cases studied, suggesting that other factors are at play in shaping binding patterns. By studying six different transcription factors, Bradley et al. were in a unique position to examine the relationships between quantitative binding divergence for different factors across the genome [22]. By performing principle component analysis (PCA) on regions bound by any factor, they found both a strong correlation between quantitative changes in binding strength across all factors (explaining 38 % of all binding divergence between *D. melanogaster* and *D. yakuba*) and both positive and negative correlations between changes in the binding of specific pairs of factors. For example, increases in binding of Gt, a repressor, were correlated with decreases in binding of Hb, an activator. A search for sequence motifs that were associated with the correlated binding divergence of all the A-P factors revealed a CAGGTAG binding motif for the zygotic transcriptional activator Vielfaltig (Vfl, also known as Zelda) [22]. This strong association between A-P factors and Vfl was later confirmed and extended into the more distant species *D. pseudoobscura* and *D. virilis* [24]. Vfl has since been shown to be a key factor in establishing regulatory regions in the early embryo that will be active later in development, and it has been suggested that it plays an important role in shaping the chromatin landscape during zygotic genome activation [56, 57]. This example highlights a case where patterns of binding con-

servation for one set of transcription factors illuminated a new functional role for a different protein as well as a general feature of *Drosophila* embryonic development.

7.6 Clustering and Combinatorial Binding

Combinatorial binding between multiple different transcription factors at the same locus, as well as homotypic clustering, in which multiple instances of the same binding motif are located close together in a regulatory element, has been proposed to play an important role in directing tissue-specific expression during *Drosophila* development and to confer robustness on regulatory outputs [7, 58]. Comparative studies of transcription factor binding can identify patterns of clustering or combinatorial binding that are conserved through evolution, providing greater evidence for their functionality, as well as assessing the degree to which clustering or its absence drives binding divergence. In their study of four A-P transcription factors, Paris et al. found that peaks in regions that were commonly bound by more than one factor were better conserved than those where only one factor bound, suggestive of a role for combinatorial binding between A-P factors [24]. The case of the two group B Sox proteins Dichaete and SoxN provides an interesting example, as these two TFs are known to bind in widely overlapping patterns and to have partially redundant roles during development [59]. Despite this apparent redundancy, Carl and Russell found that sites that are commonly bound by Dichaete and SoxN are more highly conserved than those bound by either protein alone, suggesting that such overlapping binding plays a functional role [28]. In the study of Twi in six species of *Drosophila*, the authors found that clustered Twi sites assigned to the same gene were significantly more likely to be conserved than singleton sites assigned uniquely to a gene. This effect was observed up to an inter-peak distance of 5 kb, leading the authors to suggest that Twi binding to shadow enhancers might also have an effect on ensuring robustness of gene expression patterns [23].

Observing that not all losses of Twi binding could be attributed to a corresponding loss of a Twi motif, the authors investigated whether other factors acting as binding partners for Twi had an effect on the conservation of its binding. A search for motifs that were significantly more conserved in highly conserved Twi peaks compared to divergent Twi peaks or the background genome yielded two transcription factors known to act together with Twi, Snail (Sna) and Dorsal. For Twi peaks in one species containing a Sna or Dorsal motif in addition to a conserved Twi motif, loss of the partner motif was sufficient to explain loss of Twi binding in another species in 19 % of cases. Furthermore, the top ten motifs identified in Twi binding intervals explained 49 % of losses of Twi binding despite conservation of a Twi motif. These findings go one step beyond a simple search for enriched motifs to identify those that have a functional effect on binding patterns. Integration of an evolutionary analysis of gains and losses of Twi binding with a search for conserved co-occurring motifs led to both the validation of known Twi co-regulators such as Dorsal and Sna and the identification of new factors that could potentially bind to

enhancers with Twi in a combinatorial manner to direct specific patterns of gene expression during development [23].

7.7 Insulator Function in Genome Structural Evolution

Insulator proteins are a unique category of DNA-binding proteins that have genomic functions beyond the direction of specific target gene expression, since they are believed to be involved in maintaining genomic structure and delineating domains of actively transcribed chromatin [51, 52, 55, 60, 61]. As genomes evolve on a large scale, it is expected that insulator-binding sites should evolve in tandem in order to regulate new structural and functional elements. Supporting this view, the CTCF, BEAF-32, CP190 and Mod(mdg4) insulator proteins show enriched binding near syntenic breakpoints identified in the genomes of 12 *Drosophila* species [55]. Comparative studies of these proteins can therefore highlight functional aspects of their roles in genomic evolution. In their study of CTCF binding in four species of *Drosophila*, Ni et al. reasoned that newly acquired CTCF sites in a species might be associated with new genes unique to that species [51]. After examining 42 lineage-specific genes that are essential for survival in *D. melanogaster*, they found that eight of these were associated with a newly evolved CTCF site within 5 kb. This association was significantly stronger than that observed between conserved CTCF sites and conserved essential genes, suggesting that the birth of new genes with a strong fitness effect can drive the evolution of new CTCF sites that might shape their regulation. This study also detected signatures of positive selection acting on new CTCF sites at the sequence level [51].

Another functional feature of insulator proteins in *Drosophila* is their genome-wide positioning separating divergently transcribed genes with differing patterns of expression. Enrichment for this type of positioning has been detected for BEAF-32, which was also studied in four species of *Drosophila* [52, 55]. It was found that BEAF-32 sites between divergently transcribed genes were correlated with the distance between gene pairs as well as the overall size and gene density of the genome. Thirty-two percent of gene pairs in *D. melanogaster* have a BEAF-32 site between them, whereas in *D. virilis*, which has a genome 46 % larger than that of *D. melanogaster*, the percentage decreases to 15 %. In contrast, in *D. simulans* and *D. pseudoobscura*, which have similar genome sizes and gene densities to *D. melanogaster*, the percentage of gene pairs with a BEAF-32 site (28 %) is much closer to that observed in *D. melanogaster*. A large percentage of non-conserved BEAF-32 binding sites (87 % of non-conserved sites in *D. simulans*, 41 % in *D. pseudoobscura* and 55 % in *D. virilis*) are associated with chromosomal rearrangements, which result in the rearrangement of genes with regard to their neighbours [52]. Both of these pieces of evidence support a functional role for BEAF-32 in separating the regulatory domains of neighbouring genes and allowing them to maintain separate transcriptional regimes and expression patterns.

7.8 Conclusions and Future Outlook

The hypothesis that functional regulatory elements in noncoding DNA is likely to be more conserved during evolution than the genomic background evolving at a neutral rate is a powerful one. This idea allows the leverage of comparative data to tease out the most important functional roles or features of regulatory sequences and proteins from amongst the many sources of noise inherent in genome-wide datasets. The amenability of *Drosophila* to molecular techniques and genetic manipulation, as well as the publication of the sequenced genomes and phylogenetic relationships of twelve *Drosophila* species [62] and the ongoing community efforts to sequence more species, makes the fruit fly a compelling model in which to conduct comparative studies of transcription factor binding.

Comparative studies performed with several different factors have found a range of binding divergence rates that, while broadly comparable, can differ significantly with increasing phylogenetic distance. While some of the differences may be due to the analytical methods employed, overall the findings suggest that selection may operate differently on different factors, constraining both binding levels and motifs to different extents, even when they may be located within the same regulatory domains. It appears that conservation of some factors, such as CTCF, may have a stronger correlation with conservation of the local surrounding sequence [51]; this correlation makes sense when considering the role that CTCF plays in enhancer-blocking and demarcating regulatory domains [55, 60]. In the case of classical transcription factors, comparative studies can help determine higher-confidence lists of target genes, as several studies have shown that peaks that are conserved between multiple species show enrichments for indicators of function [9, 23, 24]. Other functional aspects of regulatory architecture that have been uncovered by comparative studies in *Drosophila* include the importance of homotypic clustering of sites as well as combinatorial binding between multiple partner transcription factors and the existence of pioneer factors such as Vfl, which might play a role in shaping the chromatin landscape early in development [22, 23, 57].

It is interesting to contrast these results with those obtained in similar studies in vertebrates. In general, it appears that binding events between vertebrate species at similar phylogenetic distances are much less conserved than between *Drosophila* species [25, 53]. One notable exception to this rule, however, is the case of CTCF binding, which is more conserved in vertebrates in comparison to other factors than it is in flies. Potential explanations for these discrepancies include the vast differences in genome size and density of functional elements between vertebrates and *Drosophila* and the larger effective population size of insects in comparison to vertebrates, which tends to make natural selection more effective [25, 51, 54, 63, 64]. In the case of proteins that are deeply conserved between insects and vertebrates, such as CTCF, comparing the patterns of binding conservation in both groups can lead to insights about the shared and divergent functions of ancient orthologs as well as the evolutionary forces that have shaped each lineage.

One of the ultimate goals of studying regulatory DNA is to understand the complex chain of causality that leads from a genetic sequence to a particular phenotype. This chain must lead from transcriptional regulators through to the expression of genes in a temporally, spatially and quantitatively specific manner; as such, the degree to which quantitative variation in transcription factor binding actually influences gene expression remains an important open question. Comparisons between genome-wide RNA-seq data and transcription factor binding data give conflicting results; while expression levels and binding levels are correlated within a species, it is unclear to what extent their variation is correlated between species [24, 51]. What other factors might mediate the relationship between transcription factor binding variation and expression variation? Studies that consider the complex interactions involved in combinatorial binding and the effect of chromatin landscapes, as well as expression quantitative trait locus (eQTL) studies that leverage population data to search for loci that affect variation in gene expression, may offer some hints [65]. Moving forward, as researchers attempt to integrate many different types of genome-wide data in order to generate a comprehensive view of gene regulation, we expect that comparative studies will continue to contribute valuable insights by highlighting functional patterns that have been conserved during evolution and that *Drosophila* will remain an organism of choice for such studies, thanks to its rich collection of experimental and bioinformatics resources.

References

1. Dunham I, Kundaje A, Aldred SF et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. doi:10.1038/nature11247
2. Gordon KL, Ruvinsky I (2012) Tempo and mode in evolution of transcriptional regulation. PLoS Genet 8:e1002432. doi:10.1371/journal.pgen.1002432
3. Neph S, Vierstra J, Stergachis AB et al (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. Nature 489:83–90. doi:10.1038/nature11212
4. Wray GA (2007) The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 8:206–216. doi:10.1038/nrg2063
5. Kaplan T, Li X-Y, Sabo PJ et al (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early *Drosophila* development. PLoS Genet 7:e1001290. doi:10.1371/journal.pgen.1001290
6. Li X-Y, Thomas S, Sabo PJ et al (2011) The role of chromatin accessibility in directing the widespread, overlapping patterns of *Drosophila* transcription factor binding. Genome Biol 12:R34. doi:10.1186/gb-2011-12-4-r34
7. Zinzen RP, Girardot C, Gagneur J et al (2009) Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 462:65–70. doi:10.1038/nature08531
8. Fisher WW, Li JJ, Hammonds AS et al (2012) DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in *Drosophila*. Proc Natl Acad Sci U S A 109:21330–21335. doi:10.1073/pnas.1209589110
9. Biggin MD (2011) Animal transcription networks as highly connected, quantitative continua. Dev Cell 21:611–626. doi:10.1016/j.devcel.2011.09.008
10. MacArthur S, Li X-Y, Li J et al (2009) Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions. Genome Biol 10:R80. doi:10.1186/gb-2009-10-7-r80

11. Hare EE, Peterson BK, Iyer VN et al (2008) Sepsid even-skipped enhancers are functionally conserved in *Drosophila* despite lack of sequence conservation. PLoS Genet 4:e1000106. doi:10.1371/journal.pgen.1000106
12. Arnoult L, Su KFY, Manoel D et al (2013) Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. Science 339:1423–1426. doi:10.1126/science.1233749
13. Frankel N, Wang S, Stern DL (2012) Conserved regulatory architecture underlies parallel genetic changes and convergent phenotypic evolution. Proc Natl Acad Sci U S A 109:20975–20979. doi:10.1073/pnas.1207715109
14. Kalay G, Wittkopp PJ (2010) Nomadic enhancers: tissue-specific cis-regulatory elements of yellow have divergent genomic positions among *Drosophila* species. PLoS Genet 6:e1001222. doi:10.1371/journal.pgen.1001222
15. Greil F, Moorman C, van Steensel B (2006) DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase. Methods Enzymol 410:342–359
16. Aleksic J, Russell S (2009) ChIPing away at the genome: the new frontier travel guide. Mol BioSyst 5:1421. doi:10.1039/b906179g
17. Van Steensel B, Delrow J, Henikoff S (2001) Chromatin profiling using targeted DNA adenine methyltransferase. Nat Genet 27:304–308. doi:10.1016/S0076-6879(06)10016-6
18. The modENCODE Consortium, Roy S, Ernst J et al (2010) Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. Science 330:1787–1797. doi:10.1126/science.1198374
19. Contrino S, Smith RN, Butano D et al (2011) modMine: flexible access to modENCODE data. Nucleic Acids Res 40:D1082–D1088. doi:10.1093/nar/gkr921
20. Landt SG, Marinov GK, Kundaje A et al (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 22:1813–1831. doi:10.1101/gr.136184.111
21. Trinh QM, Jen F-YA, Zhou Z et al (2013) Cloud-based uniform ChIP-Seq processing tools for modENCODE and ENCODE. BMC Genomics 14:494. doi:10.1186/1471-2164-14-494
22. Bradley RK, Li X-Y, Trapnell C et al (2010) Binding site turnover produces pervasive quantitative changes in transcription factor binding between closely related *Drosophila* species. PLoS Biol 8:e1000343. doi:10.1371/journal.pbio.1000343
23. He Q, Bardet AF, Patton B et al (2011) High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. Nat Genet 43:414–420. doi:10.1038/ng.808
24. Paris M, Kaplan T, Li XY et al (2013) Extensive divergence of transcription factor binding in *Drosophila* embryos with highly conserved gene expression. PLoS Genet 9:e1003748. doi:10.1371/journal.pgen.1003748
25. Villar D, Flicek P, Odom DT (2014) Evolution of transcription factor binding in metazoans – mechanisms and functional implications. Nat Rev Genet 15:221–233. doi:10.1038/nrg3481
26. Choo SW, White R, Russell S (2011) Genome-wide analysis of the binding of the Hox protein Ultrabithorax and the Hox cofactor Homothorax in *Drosophila*. PLoS One 6:e14778. doi:10.1371/journal.pone.0014778
27. Vogel MJ, Peric-Hupkes D, van Steensel B (2007) Detection of in vivo protein–DNA interactions using DamID in mammalian cells. Nat Protoc 2:1467–1478. doi:10.1038/nprot.2007.148
28. Carl SH, Russell S (2015) Common binding by redundant group B Sox proteins is evolutionarily conserved in *Drosophila*. BMC Genomics 16:292. doi:10.1186/s12864-015-1495-3
29. Horn C, Wimmer EA (2000) A versatile vector set for animal transgenesis. Dev Genes Evol 210:630–637. doi:10.1007/s004270000110
30. Groth AC, Fish M, Nusse R, Calos MP (2004) Construction of transgenic *Drosophila* by using the site-specific integrase from phage φC31. Genetics 166:1775–1782. doi:10.1534/genetics.166.4.1775
31. Southall TD, Gold KS, Egger B et al (2013) Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells. Dev Cell 26:101–112. doi:10.1016/j.devcel.2013.05.020

32. Adryan B, Teichmann SA (2006) FlyTF: a systematic review of site-specific transcription factors in the fruit fly *Drosophila melanogaster*. Bioinformatics 22:1532–1533. doi:10.1093/bioinformatics/btl143
33. Hsia CC, McGinnis W (2003) Evolution of transcription factor function. Curr Opin Gen Dev 13:199–206. doi:10.1016/S0959-437X(03)00017-0
34. Kim J, He X, Sinha S (2009) Evolution of regulatory sequences in 12 *Drosophila* species. PLoS Genet 5:e1000330. doi:10.1371/journal.pgen.1000330
35. Majoros WH, Ohler U (2010) Modeling the evolution of regulatory elements by simultaneous detection and alignment with phylogenetic pair HMMs. PLoS Comput Biol 6:e1001037. doi:10.1371/journal.pcbi.1001037
36. Sinha S, He X (2007) MORPH: probabilistic alignment combined with hidden Markov models of cis-regulatory modules. PLoS Comput Biol 3:e216. doi:10.1371/journal.pcbi.0030216
37. Kheradpour P, Stark A, Roy S, Kellis M (2007) Reliable prediction of regulator targets using 12 *Drosophila* genomes. Genome Res 17:1919–1931. doi:10.1101/gr.7090407
38. Meireles-Filho A, Stark A (2009) Comparative genomics of gene regulation–conservation and divergence of cis-regulatory information. Curr Opin Genet Dev 19:565–570. doi:10.1016/j.gde.2009.10.006
39. Stark A, Lin MF, Kheradpour P et al (2007) Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature 450:219–232. doi:10.1038/nature06340
40. Halfon M, Zhu Q, Brennan E, Zhou Y (2011) Erroneous attribution of relevant transcription factor binding sites despite successful prediction of cis-regulatory modules. BMC Genomics 12:578. doi:10.1186/1471-2164-12-578
41. Sikora-Wohlfeld W, Ackermann M, Christodoulou EG et al (2013) Assessing computational methods for transcription factor target gene identification based on ChIP-seq data. PLoS Comput Biol 9:e1003342. doi:10.1371/journal.pcbi.1003342
42. Gallo SM, Gerrard DT, Miner D et al (2010) REDfly v3.0: toward a comprehensive database of transcriptional regulatory elements in *Drosophila*. Nucleic Acids Res 39:D118–D123. doi:10.1093/nar/gkq999
43. Griffith OL, Montgomery SB, Bernier B et al (2007) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36:D107–D113. doi:10.1093/nar/gkm967
44. Arnold CD, Gerlach D, Stelzer C et al (2013) Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science 339:1074–1077. doi:10.1126/science.1232542
45. Bardet AF, He Q, Zeitlinger J, Stark A (2011) A computational pipeline for comparative ChIP-seq analyses. Nat Protoc 7:45–61. doi:10.1038/nprot.2011.420
46. Fujita PA, Rhead B, Zweig AS et al (2010) The UCSC Genome Browser database: update 2011. Nucliec Acids Res 39:D876–D882. doi:10.1093/nar/gkq963
47. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. doi:10.1093/bioinformatics/btq033
48. Aszodi A (2012) MULTOVL: fast multiple overlaps of genomic regions. Bioinformatics 28:3318–3319. doi:10.1093/bioinformatics/bts607
49. Bailey T, Krajewski P, Ladunga I et al (2013) Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol 9:e1003326. doi:10.1371/journal.pcbi.1003326
50. Moses AM, Pollard DA, Nix DA et al (2006) Large-scale turnover of functional transcription factor binding sites in *Drosophila*. PLoS Comput Biol 2:e130. doi:10.1371/journal.pcbi.0020130
51. Ni X, Zhang YE, Nègre N et al (2012) Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. PLoS Biol 10:e1001420. doi:10.1371/journal.pbio.1001420
52. Yang J, Ramos E, Corces VG (2012) The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species. Genome Res 22:2199–2207. doi:10.1101/gr.142125.112
53. Schmidt D, Wilson MD, Ballester B et al (2010) Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science 328:1036–1040. doi:10.1126/science.1186176

54. Schmidt D, Schwalie PC, Wilson MD et al (2012) Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell 148:335–348. doi:10.1016/j.cell.2011.11.058
55. Nègre N, Brown CD, Shah PK et al (2010) A comprehensive map of insulator elements for the *Drosophila* genome. PLoS Genet 6:e1000814. doi:10.1371/journal.pgen.1000814
56. Harrison MM, Li X-Y, Kaplan T et al (2011) Zelda binding in the early *Drosophila* melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition. PLoS Genet 7:e1002266. doi:10.1371/journal.pgen.1002266
57. Satija R, Bradley RK (2012) The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the *Drosophila* embryo. Genome Res 22:656–665. doi:10.1101/gr.130682.111
58. Lifanov AP (2003) Homotypic regulatory clusters in *Drosophila*. Genome Res 13:579–588. doi:10.1101/gr.668403
59. Ferrero E, Fischer B, Russell S (2014) *SoxNeuro* orchestrates central nervous system specification and differentiation in *Drosophila* and is only partially redundant with *Dichaete*. Genome Biol 15:R74. doi:10.1186/gb-2014-15-5-r74
60. Mohan M, Bartkuhn M, Herold M, Philippen A (2007) The *Drosophila* insulator proteins CTCF and CP190 link enhancer blocking to body patterning. EMBO J 26:4203–4214. doi:10.1038/sj.emboj.7601851
61. Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137:1194–1211. doi:10.1016/j.cell.2009.06.001
62. Clark AG, Eisen MB, Smith DR et al (2007) Evolution of genes and genomes on the *Drosophila* phylogeny. Nature 450:203–218. doi:10.1038/nature06341
63. Kunarso G, Chia N-Y, Jeyakani J et al (2010) Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet 42:631–634. doi:10.1038/ng.600
64. Martin D, Pantoja C, Miñán AF et al (2011) Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. Nat Struct Mol Biol 18:708–714. doi:10.1038/nsmb.2059
65. Massouras A, Waszak SM, Albarca-Aguilera M et al (2012) Genomic variation and its impact on gene expression in *Drosophila melanogaster*. PLoS Genet 8:e1003055. doi:10.1371/journal.pgen.1003055
66. Hare EE, Peterson BK, Eisen MB et al (2008) A careful look at binding site reorganization in the *even-skipped* enhancers of *Drosophila* and Sepsids. PLoS Genet 4:e1000268. doi:10.1371/journal.pgen.1000268

Chapter 8
The Little Known Universe of Short Proteins in Insects: A Machine Learning Approach

Dan Ofer, Nadav Rappoport, and Michal Linial

Abstract Modern genomics and proteomics technologies are turning out immense quantities of sequenced proteins. The only feasible way to assign functions to this flood of sequences is by applying state-of-the-art computational methods for automated functional annotation. We illustrate the significance of machine learning tools in identifying and annotating short bioactive proteins and peptides from insect genomes. Over 500,000 full-length proteins from insects are currently archived in databases, of which ~15 % are short proteins. Currently, most short sequences remain uncharacterized. We developed a platform to systematically identify the functional class of short toxin-like peptides in metazoa. We present data from eight representative genomes (140,000 proteins) that cover the main phylogenetic branches of Hexapoda. The platform is a trained machine-predictor that successfully identified ~800 toxin-like candidates, 250 of them predicted with high confidence. These proteins' functions include ion channel inhibition, protease inhibitors, antimicrobial peptides, and components of the innate immune system. Our systematic approach can be expanded to new genomes and other biological classes of proteins. Using similar methodologies, we illustrate the success of identifying overlooked neuropeptide precursors. The systematic discovery of insect neuropeptides and short toxin-like proteins allows developing new strategies for pest control and manipulating insects' behavior. The overlooked secreted short peptides are discussed with respect to their evolution and potential applications in biotechnology.

D. Ofer • M. Linial (✉)
Department of Biological Chemistry, Life Sciences Institute, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: ddofer@gmail.com; michall@cc.huji.ac.il

N. Rappoport
School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel
e-mail: nadavrap@cs.huji.ac.il

C. Raman et al. (eds.), *Short Views on Insect Genomics and Proteomics*,
Entomology in Focus 3, DOI 10.1007/978-3-319-24235-4_8

Abbreviations

ANN	Artificial neural network
AUC	Area under ROC curve
CAFA	Critical automatic functional annotation
ClanTox	Classifier of animal toxins
CNS	Central nervous system
CV	Cross validation
ETH	Ecdysis-triggering hormone
HMM	Hidden Markov model
ICI	Ion channel inhibitor
MFS	Major facilitator superfamily
ML	Machine learning
MS	Mass spectrometry
nAChR	Nicotinic acetylcholine receptors
NGF	Nerve growth factor
NP	Neuropeptide
NPP	Neuropeptide precursor
OCLP	ω-Conotoxin-like protein
PSSM	Position-specific scoring matrix
SP	Signal peptide
SVM	Support vector machine
TIL	Trypsin inhibitor like
TOLIPs	Toxin-like proteins
TLS	Toxin-like stability

8.1 Automated Functional Classification of Proteins: Sequence Similarity

In recent years, there has been an exponential increase in biological data, particularly of gene and protein sequences. The rapid growth rate of protein sequence data cannot be handled by performing individual experimental studies to determine the function(s) of every single protein, as was traditionally the case. Therefore, computational prediction is currently the only feasible approach for high-throughput identification of protein function [1].

Generally, functional classification is performed using a supervised approach, i.e., inferring functional classification for a sequence according to existing sequences whose functions are known. The most naïve, supervised approach is the *nearest-neighbor* search [2]. In practical terms, a database of sequences is searched for a query sequence with the goal of identifying similar sequences. The most common algorithms and search engines that perform this task are BLAST and FASTA [3]. If

a significantly similar sequence is found, the query sequence will be considered to possess a similar function; this concept is often considered as "guilt by association." The "rule of the thumb" for this inference has been defined as the *twilight zone* concept [4]: for a sequence at least 100 amino acids long, it is most likely to be a homologue if at least 30 % of the amino acids are identical. Below this value, the sequence is in the "twilight zone," where the similarity cannot be separated from randomly occurring similarity.

Although this direct inference approach is useful for many sequences, it suffers from critical caveats:

1. In order to learn about a sequence, there must exist a significantly similar sequence whose function is known, essentially precluding function prediction for unknown protein families.
2. Many proteins with similar sequences have different functions and would therefore be mistakenly classified as having the same function. Such cases are common for *paralogs*.
3. Many proteins exist that share functionalities and active sites or domains but possess significantly different sequences, despite having similar functions.

8.2 Functional Classification of Proteins: An Ill-Defined Term

There is an obvious connection between the "granularity" of a function (general or specific) and the evolutionary diversity of the proteins that share it [5]. Typically, groups of proteins that share a high-level functionality (i.e., enzymes) are much more diverse than low-level (e.g., urease enzymes) functionality groups [6]. This simple notion serves to define a scoring method for functional similarities [7]. The critical automatic functional annotation (abbreviated CAFA) initiative serves to set a measure for the success and failure in functional assignment. Open competitions for functional assignment over thousands of proteins show that there is considerable room for improvement [8].

There is an obvious interest in having classification machines successfully learn *high-level functionality*. If the training set consists of proteins that share a low-level functionality, the classifier would only be able to detect proteins that belong to the narrow function that was learned. Essentially, this would reproduce the main caveat of the nearest-neighbor search, i.e., the need to have a known, near-identical representative of every possible biological functional group. However, if the training set consists of proteins that share a high-level function, the classifier will be able to detect any protein that belongs to a very broad class, even if "close" representatives are unknown.

The point can best be demonstrated via an example: Consider the case of the major facilitator superfamily (MFS). This superfamily includes over 300,000 proteins

capable of transporting small solutes in response to ion gradients [9]. In general terms, proteins of the family belong to the transmembrane transport system. However, if we use classifiers of low-level functionality, we would have a set of classifiers that classify the different types of MSFs (including polyol permease, nitrate transporter, multidrug resistance protein, sialic acid transporter, and many more). If a sequence of a novel subtype of MSF were found, we would not be able to identify its function at the lower level as it would not belong to any known transporter family. However, if we had access to an MSF classifier, we could identify the sequence as a novel type of MSF transporter. An illustration of such an instance is shown in Fig. 8.1.

It is difficult to learn high-level functionality computationally, particularly when we might barely understand them on the theoretical level. While we might expect nitrate transporters from different organisms to share similar sequences due to evolutionary homology, we would not expect this of a high-level group such as all transmembrane transporters. "Statistical modeling" methodologies can then be applied [10]. Central to the success of these methods is construction of multiple sequence alignments [11]. It is plausible to characterize a new sequence with direct sequence-based techniques such as position-specific scoring matrices (PSSMs) and hidden Markov models (HMMs) [12]. These methods are widely used for sequence classification and identification. A large resource for families and domains in proteins is structural knowledge (often reliant on experimental sources), together with statistical modeling of each family [12]. A resource that relies entirely on automatic learning schemes provides a complementary view [13]. These methods were successfully applied to characterize features such as conserved positions and active

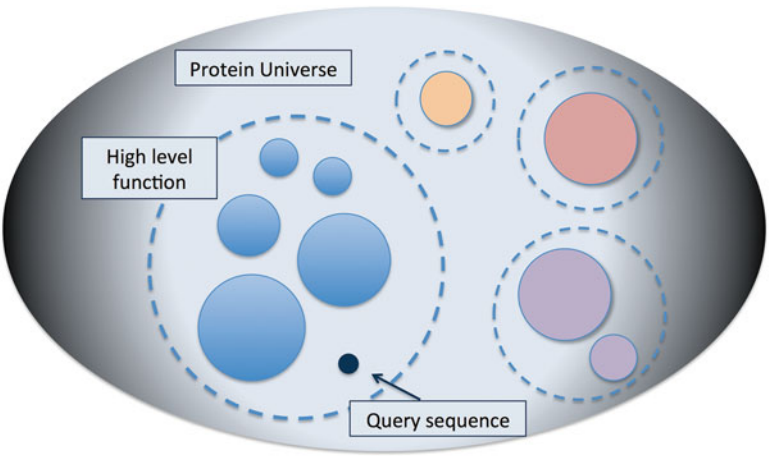


Fig. 8.1 Assigning functional annotation to a novel protein. Functional classes are represented as *colored circles* (i.e., the protein sequences that are similar and functionally related appear next to each other in this space). The level of functionality is relative to the diameter of the *dashed circles*. The *black dot* represents a novel sequence of unknown function. In this example, the sequence does not belong to any low-level functional classes (marked as *circles*) but does belong to a high-level function class (the *dashed circle*)

sites in enzymes and other binding sites in proteins [14]. However, there are cases in which multiple sequence alignment is unfeasible or simply uninformative.

8.3 Functional Classification of Proteins: Machine Learning

When sequence alignment is meaningless (e.g., only a small number of sequences can be aligned, or the information content of the multiple alignment is minimal), other methodologies have to be adapted. Some methods produce function classifiers that are not directly sequence based, but rather rely on "global" sequence-derived quantitative features that are extracted from the sequence and typically do not take amino acid sequential positions into consideration (e.g., the length of the protein, the amino acid frequencies, the weight of the protein) [15]. In such instances, the sequences are transformed into multidimensional space where each protein is represented by vectors of features in that space. Then, a classifier is learned by a statistical approach such as a support vector machine (SVM) [16] or a decision tree classifier method such as random forests [17]. Such methods which avoid the requirement for sequence alignments have shown success in learning high-level functional traits (such as the high level of protein family structural folds) while often being far more computationally efficient. While both the direct sequence-based approaches and the sequence-derived feature approaches may use the same information as input, namely, the sequence itself, they can perform very differently due to the manner in which they exploit the data and the information they extract from it. There are cases in which an intelligent choice of numerical features (i.e., those that can best capture the characteristics of the relevant sequences) can significantly outperform popular alignment models (e.g., HMM and PSSM) and can be extended for a variety of other data structures such as gene co-expression data.

We have now set the stage and background on the difficulties and solutions for functional inference of novel protein sequences. In this chapter, we show how for large sets of insect proteins we applied statistical learning methods to annotate unknown sequences as belonging to distinct, high-level functional classes. We considered sequences that are impossible to characterize by existing direct sequence-based methods (because the sequences are not alignable), but for which global, sequence-derived features can successfully characterize them exceptionally well.

The statistical learning technique framework is based on the notion of supervised machine learning methods, in which a group of known, annotated sequences serve for the *learning/training* phase. This group of sequences is referred to as the *training set*. Once the learning stage is complete, the characterization that was computationally learned, known as the *hypothesis*, can be used to classify unidentified sequences. The advantage of machine learning methods over the sequence similarity, nearest-neighbor approach (described above) is in the fact that the learning methods can identify the minimal conserved set of characteristics in each family of proteins and focus on searching only for these characteristics. This makes the learning methods much more powerful than the naïve sequence similarity approach.

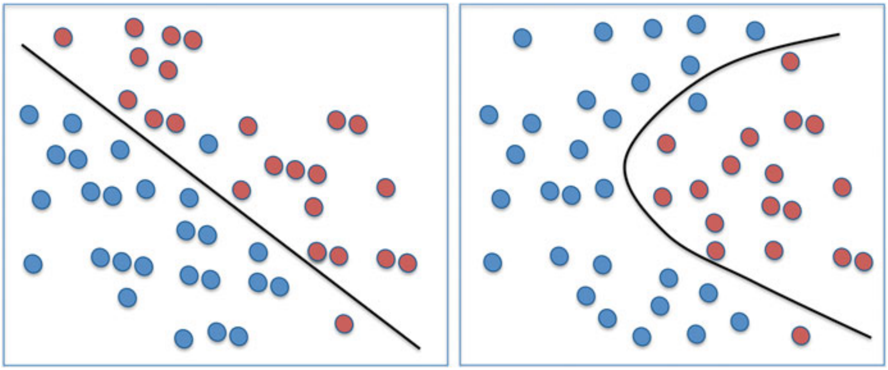


Fig. 8.2 Support vector machine (SVM) classification between labeled instances (*red* and *blue*). The *black line* represents the separator. The scenarios show a linear and a nonlinear SVM classification

The principal goal in the machine learning approach is to regard the problem as one of supervised classification or prediction in a binary (i.e., "Yes/No," "True/False," Positive/Negative) or a multi-class prediction problem. In a binary classification problem setting (such as SVM classification) [18], the goal is to classify two classes of points (indicated as positives and negatives) by constructing an optimal separator according to distinguishing features of items belonging to those sets so as to distinguish optimally between them and to classify new instances as belonging to one class or the other. The separator is set to ensure a maximal margin to the points (Fig. 8.2). The separator does not have to be linear and derives from a class of similarity measures (so-called kernel functions). The use of such a technique provides a set of "classifiers" that can then be tested and then used to predict "unlabeled" new instances [19].

The field of machine learning is immense with a strong impact on predictive biology [20, 21]. Machine learning technologies cover supervised methods of binary classification including various SVMs and kernels, decision trees, artificial neural networks (ANNs), ensembles of classifiers (such as random forests and AdaBoost) [17], and unsupervised methods for clustering [22].

8.4 Short Proteins: An Overlooked Niche

The ability to learn about a protein by comparing it to its (inferred) homologues has been used in functional prediction, secondary structure prediction, three-dimensional fold prediction, and several other applications. However, the power of sequence similarity-based tools is greatly diminished for short protein sequences. This is because when comparing short sequences it is difficult to distinguish genuine homology from mere evolutionary noise/coincidence. For example, submitting a short amino acid sequence to a sequence similarity search server such as BLAST

will usually result in matches with barely significant or insignificant e-values (a statistical measure of significance for the expectation value), even for sequences with high percentages of identity. Therefore, the detection of homologues for short proteins by using sequence similarity tends to fail.

The difficulty in the identification of homologues is only one of the problems associated with short proteins [23]. Let us consider newly sequenced genomes. The first task is identifying potential (putative) gene products. The main steps for identifying the encoded proteins include:

1. Sequence similarity: While this method is the most powerful computational approach, as indicated, it fails to detect short proteins.
2. Comparative genomics: This method requires the aligned genomes of related species. Additionally, this method is likely to fail to detect short proteins for similar reasons to the sequence similarity approach.
3. Ab initio gene prediction: The default parameters require a minimal length for potential ORFs (open reading frames), which may further hinder the detection of short proteins.
4. High coverage of the transcriptome and proteome by high-throughput experimental technologies.

Some experimental methods focus on detection of mRNA expression and others on detection of protein expression [24]. High-throughput experiments for the detection of evidence for mRNA expression are perhaps the best source of data for detecting new proteins but are often far from comprehensive due to the fact that many genes are only expressed under certain conditions and technical reasons [25]. Deep sequencing technologies (e.g., RNA-Seq of the transcriptome) can overcome the problem of low coverage. However, a flood of short noncoding and fragmented transcripts are also detected with the potential to mask the transcripts of short proteins. The most direct approach is mass spectrometry (MS) proteomics. Nevertheless, this high-throughput protein expression technology requires special tweaking in order to detect short proteins and is limited to the detection of highly expressed proteins. Furthermore, if a short protein is not already a known candidate, it will not be found [26]. As a consequence of these computational and experimental difficulties, it is conceivable that short proteins represent a relatively understudied and neglected niche [23].

8.5 Short Proteins: Why Do We Care?

Considering the identification of short proteins as a possibly underrepresented group, one might pause to ask how many short proteins are there and what kinds of functions can be associated with them. Examination of the SwissProt database [27] shows that 2 % of the registered sequences (excluding fragments) are less than 50 amino acids in length and 10 % are shorter than 100 amino acids. Clearly, the cellular machinery is capable of producing many functioning small proteins, and these proteins are involved in biological activities.

What biological functions do these proteins fulfill? To answer this question, we performed a simple statistical enrichment test for biological functions on the group of all SwissProt proteins shorter than 100 amino acids [28]. The enrichment test is aimed at finding biological groups that appear significantly more often than expected for a random selection of proteins. Several biological groups and functions are highly overrepresented among short proteins. The significance of being short for neuropeptides (NPs) is evident; of 1430 annotated proteins, 1170 were shorter than 100 amino acids [29]. Other enriched keywords include "signal peptide," "secretion," and more. The functional groups that had the highest statistical enrichment value include "toxin," "neurotoxin," "ion channel inhibitor" (ICI), "sodium channel inhibitor," and "scorpion long-chain toxin." For example, from the 2080 proteins that are annotated as being ICIs, 1900 are less than 100 amino acids long. This group includes most animal toxins.

8.5.1 The ID of Animal Toxins

Toxins are animal venom proteins aimed at inflicting harm to the organism on which the venom acts. They are extremely varied in terms of function and effect and include ion channel inhibitors (ICIs), phospholipases, protease inhibitors, disintegrins, membrane pore inducers, and more [30].

ICIs constitute the most widely studied group of toxins. Even specific groups of ICIs which inhibit the same channel type are varied in sequence and structural folds [31]. One group of ICIs whose evolution has been previously studied is the potassium ion channel inhibitors (K^+ ICIs). K^+ ICIs are found in a wide variety of venomous species and possess at least ten different structural folds [32]. In spite of this, all K^+ ICIs possess two residues that are critical for function, a Lys and a Tyr or Phe, which are known as the functional dyad [33]. Surprisingly, even though these residues appear in very different positions in the sequences of K^+ ICIs, the solved structures show they are closely aligned in space relative to each other.

On the other hand, some scorpion toxins, while sharing the same structural fold, act to inhibit different ion channels, including Ca^{2+}, K^+, Na^+, and Cl^-. This surprising observation shows that although there are many toxin folds, none is definitively associated with any particular ion channel selectivity [31]. This raises interesting questions regarding evolutionary convergence, divergence, and functional conservation.

8.5.2 TOLIPs: Endogenous Toxin-Like Proteins

Toxins appear only in very specific branches of the evolutionary tree. However, these branches are widely dispersed, including insects, snakes, sea anemones, spiders, the marine cone snail, and even mammals. Still, many toxins possessing similar functions (e.g., ICIs) appear in several unrelated venomous species. The

possibility to detect endogenous toxin-like proteins (called TOLIPs) is attractive for a number of reasons. For example, maintaining their function as ion channel blockers provides a new layer of regulation at the protein activity level (i.e., blocking ion flux through the channel). Additionally, an extensive search through the literature shows that toxin-like proteins exist in multiple species and are expressed in a wide range of nonvenomous organisms and tissues. Two striking examples are LYNX1-Ly6 [34] and SLURP-1 [35] from mammals. These are human proteins that not only possess similarity to snake α-neurotoxins but also modulate nicotinic acetylcholine receptors (nAChR) as do α-neurotoxins. The identification of SLURP-1 as an epidermal neuromodulator has helped explain the phenotype of the Mal de Meleda disease, a skin disease that results from improper activation of TNF-α [35]. A crucial question that arises in the field of insect genomics and proteomics is how many of the potential TOLIPs have actually been discovered.

8.5.3 Neuropeptides: Master Regulators of Insect Life

NPs regulate most aspects of insect life, from growth to behavior. The effect of NPs can only be fully appreciated by taking in the complementary view of their receptors and the underlying signaling network [36]. Due to the central role of insect NPs in mating behavior, growth, and reproduction, they are attractive targets for management of pests in agriculture. In our study, we addressed NPs as short neuromodulatory peptides that possess fundamental physiological roles [29, 37].

NPs are key modulators in behavior, sensation, and homeostasis [38]. Similar to endogenous TOLIPs, these peptides function in biological communication for a wide range of metazoans, from cnidarians to bilaterians, including mammals. The NPs are very short active peptides (5–30 amino acids) produced from parts of longer precursor molecules that are subjected to multiple cleavages. The posttranslational end products are subsequently modified and secreted. It is estimated that there are tens of NP precursor (NPP) genes in *Drosophila* and the honeybee. This rough estimation is based on annotations derived from only a small number of model organisms [39]. Similar to TOLIPs, the NP sequences are mostly non-alignable. Sequence similarity methods fail to predict or provide a comprehensive catalogue of NP bioactive peptides or their precursors. Active NPs in insects are diverse in terms of the site of action, the pattern of modification, and the specificity [40]. For example, the same NP may act both in the central nervous system (CNS) and as a hormone in the hemolymph, leading to different physiological responses.

8.5.4 From Features to Predictors

The goal of this section is to present a systematic approach for identifying insect TOLIPs as well as candidate NPs. We provide the analysis for a large number of insect proteomes that are archived in insect genomic resources and in central

resources such as the UniProtKB protein database. Many NPs play roles in regulating the behavior and physiology of larger animals, notably in terms of metabolism, pain regulation, and social behavior. Generating a catalogue of the proteome's short bioactive peptides (i.e., functional peptidome) will benefit the biotechnological community that seeks new directions for pest management and manipulating insect behavior in general.

We set out to construct two machine learning classifiers that are trained on: (1) animal ICI toxins and (2) NPs and NPP genes. For both types of short protein active modulators, characterization by sequence alignment-based methods is ineffective. Hence, the main logic in our approach is to identify the features that capture the characteristics of the types of proteins we seek to identify.

The scheme we present is composed of three main parts: (1) data analysis from genomes to short proteins, (2) the (supervised) machine learning approach and prediction, and (3) the annotation and functional validation phase.

The workflow for data mining and prediction using machine learning is composed of several steps: (1) acquiring the appropriate data in the form of protein sequences of the desired class and selecting "negative" sequences; (2) extracting features derived from the selected sequences; (3) constructing the training sets, training the classifier(s) according to the data and the generated features, and validating the predictions; (4) testing the classifier and comprehending its predictions; and (5) applying the classifiers to new proteins and selecting top predictions for further validation.

The result of all this is the discovery and subsequent annotation of new TOLIPs and NPPs (Fig. 8.3). While we discuss here the application in the context of insects, the protocol is applicable to any genome or proteome.

We will illustrate the protocol for the case of the animal toxin classifier. The ICIs share a general characteristic that can be described as structural stability (in short, toxin-like stability, TLS). Importantly, in most instances, the TLS is governed by the presence of disulfide bridges that are formed from cysteine residues along the sequence in question. The apparent rigidity of the scaffold of the proteins, together with posttranslational modifications (e.g., glycosylation), imposes rigid structural constraints.

One of the keys for a successful prediction when using machine learning is the selection of those features that may best characterize the targets we wish to predict (namely, to best separate TOLIPs from non-TOLIPs). The choice of features here was guided by the notion of stability, which is known to be associated with a large number of disulfide bridges. Therefore, the features were constructed so that they could reflect cysteine-mediated stability by encoding properties such as the frequency and the spread of cysteine residues within the sequence. In the predictor, called ClanTox (classifier of animal toxins), we used 545 features, many of which captured the TLS [41]. However, these were not restricted to cysteine-related features and were applied to all amino acids and other structural and sequence-derived qualities. The method used here represents each sequence as a vector that contains various numerical sequence features.

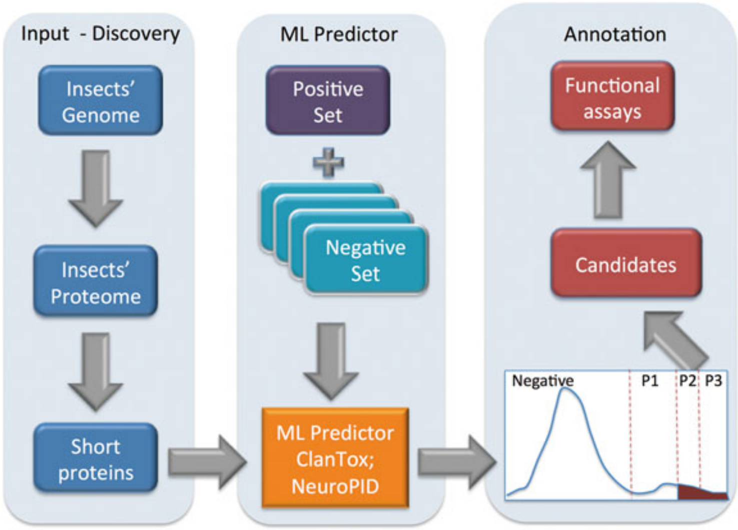


Fig. 8.3 The flow from an unannotated insect genome to the discovery of functional bioactive peptides. The scheme shows the training of the machine learning tools ClanTox and NeuroPID. These are two prediction platforms for the discovery of TOLIPs and NPs+NPPs

The general properties we examined that applied to all proteins included amino acid frequencies (20 features), amino acid pair frequencies (400 features), and sequence length (1 feature), among others. These features were shared between the two predictors (for TOLIPs and for NPPs). We will not discuss the selection and removal of redundancy from the training set, compiling alternative negative sets, tuning of the machine learning model parameters, or the cross-validation protocol.

UniProtKB and specifically the SwissProt database remain reliable sources of annotated sequences of complete proteomes and also for collecting information on toxins [42]. Eight thousand seven hundred insect proteins from UniProtKB were used as a background for testing the ClanTox predictor [41]. The results of testing short insect proteins (length <120 amino acids) from the SwissProt database (used as the input) are summarized in Table 8.1.

Using only the SwissProt database, we identified ~270 proteins ranked as putative TOLIPs, ~60 of them at a high level of predictive confidence. For example, the prediction from *Bombyx mori* includes a large number of bombyxins (types B, C, D, and G), chorion class high-cysteine proteins, fungal-chymotrypsin-trypsin inhibitors, and eclosion hormone. While each of these proteins is activated under different stimuli and at a specific developmental stage, there is only a limited set of functions that are enriched among the top-scoring predictions (Table 8.1). These functions include signaling of the innate immune system, serine protease inhibitors, and

Table 8.1 Statistically enriched keywords among the TOLIP predictions from SwissProt short proteins from insects

SwissProt keywords	Enrichment (Bonferroni)[a]
Disulfide bond	1.0E-55
Defensin	8.0E-16
Secreted	4.9E-08
Ion channel inhibitor	1.1E-07
Signal	1.1E-06
Cleavage on pair of basic residues	1.3E-06
Neurotoxin	2.4E-05
Protease inhibitor	9.0E-05
Serine protease inhibitor	9.0E-05
Toxin	6.7E-04
Knottin	7.7E-04
Hormone	3.2E-03
Zinc	1.0E-02
Antimicrobial	1.8E-02
Fungicide	2.1E-02
Calcium channel inhibitor	3.7E-02
Metal binding	3.9E-02

[a]Enrichment of keywords (p-value <0.05, Bonferroni correction), with all insect sequences of length <120 amino acids as the background set

antimicrobial functions. Considering that the training process was performed only on ICIs, it is remarkable to note that high-confidence TOLIPs share modulatory and signaling functions in development (e.g., bombyxins), the immune system (e.g., defensins, antimicrobial), and modulating tissues (e.g., protease inhibitors).

8.6 Test Case: TOLIPs in the Curated *D. melanogaster* Genome

From a set of thousands of sequences, we seek the predictor to announce for each sequence whether or not it is a toxin (or TOLIP). Discovery of overlooked TOLIPs calls for validating the top predictions (Fig. 8.3).

The most studied insect, *Drosophila melanogaster*, serves as a "testing ground." We applied the prediction platform to the complete proteome (almost 20,000 annotated genes in UniProtKB). One hundred sixty-one proteins were predicted to be TOLIPs by the ClanTox platform, with half of them ranked at the top confidence scale. Despite the high level of curation for *D. melanogaster*, about 60 % of the predictions were uncharacterized. However, most of these TOLIPs carry the signature of protease inhibitors (e.g., Kazal domain). Apparently, some TOLIPs with

Kazal domains have antibacterial and antifungal activities [43]. The rest are proteins belonging to drosomycins, sperm and seminal fluid, and metallothioneins [44].

The drosomycin family (seven sequences positively predicted) demonstrates the diversity of TOLIPs. Drosomycins are short, secreted proteins that possess antifungal activity. Similar to classical toxin ICIs, after removal of the signal peptide (SP), the mature peptides circulate in the hemolymph [45]. There, as part of the innate immune system, they act as ligands to alter intracellular signaling pathways.

8.7 Overlooked TOLIPs in Honeybee

The honeybee (*Apis mellifera*), the first sequenced genome of a venomous insect, was used for testing such a discovery phase, as its annotation level is only partial.

Among the 66 positive predictions, 26 suggest a higher level of confidence for being TOLIPs. Among them 73 % are named "uncharacterized." We observed that almost all possess a signal peptide (cleavable 20–25 amino acids at the N-terminal) [46]. About 50 % of the predicted proteins share a trypsin inhibitor-like (TIL) domain. The predictions with the highest scores are listed in Table 8.2. Structural representatives are shown in Fig. 8.4.

Once a top candidate TOLIP is detected, traditional state-of-the-art (albeit limited) sequence similarity methods can be activated. As illustrated for the sequence of H9KQJ7, applying sensitive tools for detecting remote homologues revealed a rich and a surprising resemblance to ω-conotoxins and to a set of related sequences. The multiple sequence alignment (Fig. 8.5) shows conservation of the several

Table 8.2 A sample of the top predictions of TOLIPs from *Apis mellifera*

Entry	Protein names	Len[a]	SP[b]	Family/PDB model	New discovery
P56587	Tertiapin (TPN)	21	Sec	PDB: 1TER	Tertiapin toxin like
H9K243	Uncharacterized	29	Frag	PDB: XU1	TNF and conotoxin
P01500	Apamin	46	Yes	PDB: 1TER	Tertiapin toxin like
B7UUK0	Apamin protein	46	Yes	PDB: 1TER	Tertiapin toxin like
H9KCD7	Uncharacterized	47	Yes	IPR: Zn-Fg	Zinc finger
P01499	Degranulating	50	Yes	PDB: 1TER	Tertiapin toxin like
H9K853	Uncharacterized	50	Yes	PDB: 1TER	Tertiapin toxin like
H9K3H8	Uncharacterized	58	Yes	PDB:1HP8	p8MTCP1 oncogene
P83563	Allergen Api m 6	71	Sec	PDB:1CCV	Trypsin inhibitor like
H9KEA0	Uncharacterized	71	Yes	TIL	Trypsin inhibitor like
H9KQJ7	Uncharacterized	74	Yes	PDB: 2WH9	ICI, OCLP, ω-conotoxin
Q27SJ8	Allergen Api m 6-1	92	Yes	TIL	Trypsin inhibitor like

Sec secreted, *Frag* fragment
[a]*Len* length in amino acids
[b]*SP* signal peptide

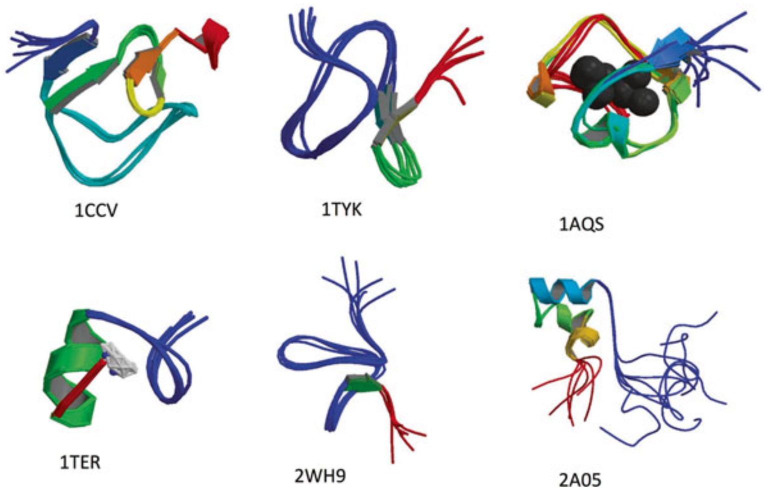


Fig. 8.4 Representatives of the top predictions for TOLIPs. The PDB accession ID is shown. The 3D structures were determined by NMR. *Connecting lines* indicate disulfide bonds; copper ions are shown as *black balls*. Sequences are colored (*rainbow*) from N- to C-terminals. For details, see text and Table 8.2

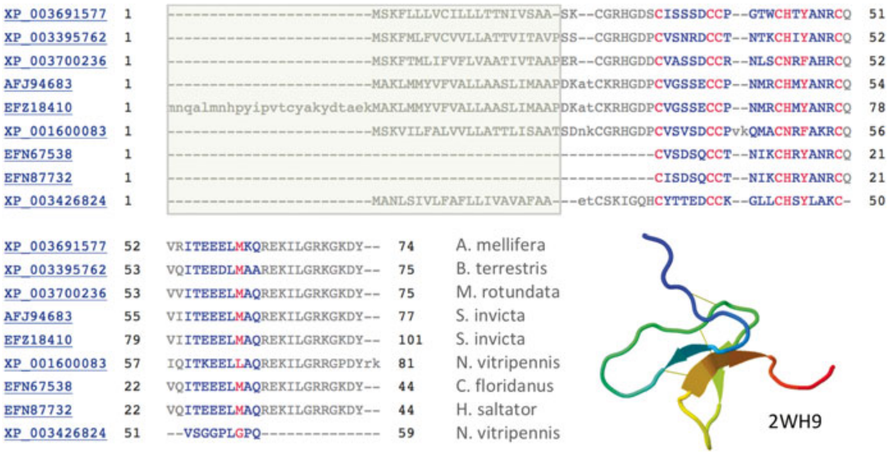


Fig. 8.5 Multiple sequence alignment of H9KQJ7 from *Apis mellifera*. The most similar sequences are from Hymenoptera including several bees, a wasp, and a number of ants. The most conserved amino acids are shown in *red*. These sequences are best modeled to PDB: 2WH9. This 3D prototype is a neurotoxin which was isolated from *Plesiophrictus guangxiensis* (tarantula) venom. The active peptide inhibits the Kv2.1 channel in human pancreatic β-cells

cysteines, which are critical in terms of structure. In addition, some positions are also conserved. Most notable is the Phe/Tyr in position 46 (F/Y numbered by the full-length H9KQJ7 sequence). It is known that the aromatic F/Y followed by N/A/R comprises the key amino acids for binding specificity to several ion channels

[47]. From the structural perspective, H9KQJ7 resembles a classical ICI from tarantula (PDB: 2WH9) as well as a large collection of structurally solved ICIs including omega-conotoxin, jingzhaotoxin, hanatoxin, huwentoxin-I, hainantoxin-I, and heteropodatoxin. Interestingly, a reversible effect of the honeybee H9KQJ7 protein on Ca^{2+} channel activity has been confirmed experimentally [48].

Several cDNAs provide supporting evidence for the expression pattern of such ω-conotoxin-like proteins (called OCLP, omega-conotoxin-like protein). OCLP-related cDNAs are found in *Anopheles gambiae*, *A. funestus*, *Aedes aegypti*, *D. melanogaster*, *Manduca sexta*, and *Heliconius erato*. None of these sequences had been previously characterized as ICI.

8.8 Evolutionary Diversity of ω-Conotoxin-Like Proteins in Insects

OCLP from honeybee has strong support for being a TOLIP: (1) It possesses a signal peptide; (2) it shares sequence similarity with assassin bug voltage-gated Ca^{2+} ICIs; (3) and structural modeling assigned the sequence to ω-conotoxin and related toxins with very high confidence (see Fig. 8.5). The expression of OCLP is exclusive to the brain (Linial and Bloch, unpublished).

A remote homologue search identified proteins in *A. gambiae* and *Ae. aegypti* containing multiple units of the OCL (omega-conotoxin-like) domain. Such organization is actually the hallmark of neuropeptides but was also noted for toxins (e.g., sarafotoxin; [49]). Remarkably, other toxins and functional motifs share the core of the OCL motif, specifically, covalitoxin II from tarantula and POI (phenol oxidase inhibitor) from *Musca domestica* [48]. The OCLP in honeybee is similar to these toxins but also to Ptu1 and ADO1, two related toxins from the assassin bugs *Peirates turpis* and *Agriosphodrus dohrni*, respectively. The function of Ptu1 as an effective Ca^{2+} channel blocker has been confirmed [50]. The OCL domain is conserved also in the freshwater planarian (*Schmidtea mediterranea)* sequence [48]. Focusing on the expansion of OCL domain in insects reveals duplication events in distinct branches along the insects' phylogeny (Fig. 8.6). There are nine OCL domains from *Nasonia vitripennis* that appear in five proteins. The repeated nature of OCL domains occurs also in proteins from *A. gambiae* and *Ae. Aegypti* (Fig. 8.6).

The short proteins discussed here raise the question of the evolutionary origin of proteins that share the OCL domain. The evolutionary relatedness that combines insects, the cone snail, and the flatworm planaria strongly argues for the importance of this fold in a diverse ecological setting. The possibility of de novo evolution for short proteins has been presented [51] and was supported by tracing the recent expansion of short immune-related proteins in mammals. Although evolutionary divergence is the most plausible explanation, convergent evolution for toxin-like proteins cannot be excluded.

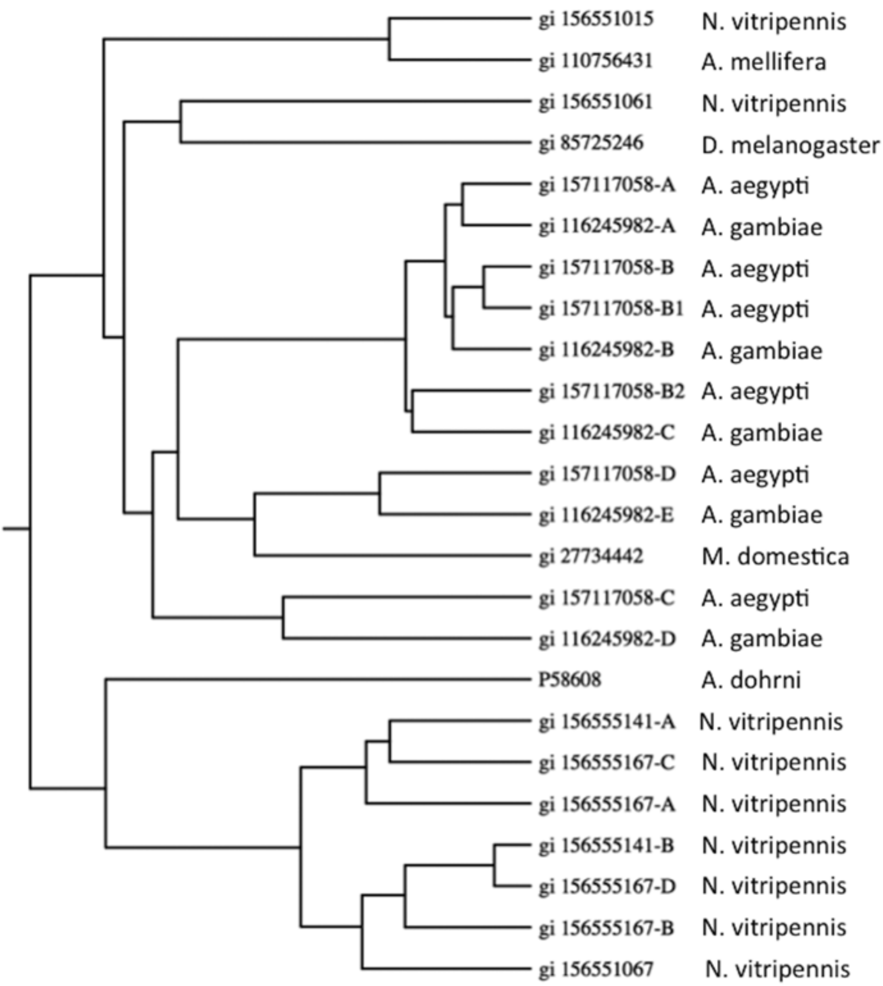


Fig. 8.6 Homology distance tree of insect proteins that contain the OCL domain. OCLP from the honeybee and a collection of other insects are shown. The OCL domains share an identical structure to ω-conotoxins with three cysteine bridges that govern the stable and compact structure of the OCL domains. The protein identifier is based on NCBI protein database

8.9 Overlooked TOLIPs in Fully Sequenced Insect Genomes

Application of the same protocol applied for the honeybee (Table 8.2) to all other insects whose genomes have been completed reveals that hundreds of overlooked TOLIPs can be traced across the entire phylogenetic tree. It is important to note that the sequences of the honeybee were not included in the training set for prediction of honeybee proteins. The same is true for the other recently sequenced genomes which were practically unavailable when the ClanTox predictor was trained. The

high-confidence predictions for these genomes reach a total of 235 (from 790 positive predictions). The most dominant functions associated with ClanTox predictions are the trypsin/chymotrypsin inhibitors, metallothioneins, TGF like, growth factor domains, defensin like, ICIs, and membrane-disrupting peptides.

Acyrthosiphon pisum (pea aphid) is represented by almost 40,000 protein sequences. The many short proteins (~9500) reflect the high number of fragmented sequences. When all the short proteins were tested using the ClanTox platform, only 18 sequences were predicted at a high level of confidence; another 112 sequences were predicted with only moderate confidence. One short sequence (J9KHE3, 63 amino acids) is a secreted protein that resembles a cysteine-rich secretory protein domain of Tpx-1 which is related to ion channel toxins and regulates ryanodine receptor Ca^{2+} signaling (PDB: 2A05, Fig. 8.4). The rest of the proteins from pea aphid are unlikely to act as secreted cell modulators.

The complete genome of the silk moth *B. mori* provides a glimpse of a different branch of the insect tree. There, several neurohormone proteins, including the insulin-like bombyxins, were positively predicted as being TOLIPs. The main functions of bombyxins are as growth factors for wing imaginal disks [52] and for general promotion and regulation of growth and metabolism [53]. The *B. mori* protein H9JHN8 is another example of a secreted protein which is uncharacterized and captures the characteristics of an overlooked TOLIP. A structural view identified numerous proteins that resemble the following functions: (1) spaetzle protein from *Drosophila*, which acts in development and in the immune system; (2) a protein from horseshoe crab involved in hemostasis and host defense; and (3) classical neurotrophins including β-nerve growth factor, brain-derived neurotrophic factor, and neurotrophin 3/4.

A surprisingly high number of positively predicted TOLIPs were associated with *A. gambiae*. There were 51 high-confidence predicted TOLIPs, many characterized by a trypsin inhibitor-like (TIL) domain (Table 8.2). Several other domains were also detected including EGF like, WAP (whey acidic protein), and elafin. A representative of the elafin family is a short secreted protein (Q7Q332) that resembles a large number of snake toxin proteins. The similarity to the 3D structure of nawaprin (PDB: 1UDK) from the venom of the spitting cobra, *Naja nigricollis*, is striking. The nonconventional circular structure is stabilized by the presence of four disulfide bonds. Interestingly, the nawaprin and elafin proteins (represented by the human leukocyte elastase-specific inhibitor) share several unique structural features but minimal sequence similarity. The functions of nawaprin or Q7Q332 from the mosquito are still not known.

Rhodnius prolixus is the most important vector of the Chagas parasite in Africa [54]. The complete genome was determined, but it is poorly annotated. Overlooked TOLIP detection using ClanTox identified T1H9H6. The sequence resembles a three-fingered fold that is abundant in snake venoms including cardiotoxins, denmotoxin, and α-bungarotoxin. These sequences belong to the diverse family named Upar/Ly6 [55]; many of the proteins are membrane markers of cells that belong to the innate immune system. The protein resemblance to snake neurotoxin (e.g., cobra) has been reported [56].

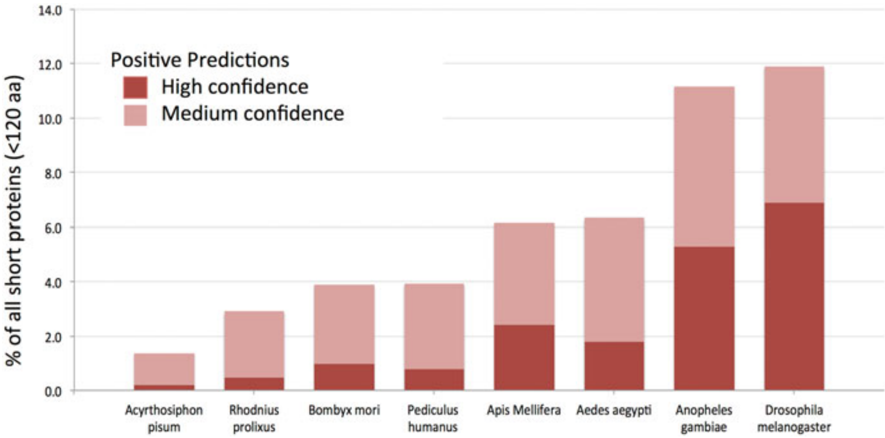


Fig. 8.7 A histogram showing the fraction (in %) of positive predictions with respect to short proteins in the indicated genomes. Several insect representatives from complete proteomes are listed. The number of high-confidence predictions ranges from 9 to 66 for *Pediculus humanus* and *Drosophila melanogaster*, respectively

In sum, the fraction of predicted TOLIPs among the short proteins varies drastically among insect genomes (Fig. 8.7). We attribute these large differences to the quality of the genome assembly. From a biological perspective, it may reflect varying complexities in the modulation of cell communication, the immune system, and/or neuronal functions.

The discovery of TOLIPs in insects led to an unexpected finding that showed the abundance of TOLIPs in viruses [57]. A cross talk of insects and their viruses was proposed. For example, protein B6S6X8 (113 aa) from *Betabaculovirus* is similar to many of the short peptides in *Drosophila* proteomes [57]. In another instance, a cysteine-rich encoding region was transferred from the endoparasitic wasp *Campoletis sonorensis* to a symbiotic polydnavirus (CsPDV) [58].

8.10 Neuropeptide Precursors in Insects

The ideas that exemplified TOLIPs and the methods used were successfully applied to identify neuropeptide precursor (NPP) genes. Neuropeptides are the products of a posttranslational regulated process of cleavage and modification from NPPs. The mature peptides are secreted from neurons and thus are collectively called neuropeptides (NPs). NPs act through their direct interaction with their receptors on presynaptic or postsynaptic cells [59]. In insects, NPs function in cell communication and affect social behaviors, including mating, food uptake, and metabolism [60].

Insects have evolved a large repertoire of NPs. Figure 8.8 (left) shows the number of annotated NPs from major taxonomical groups. We considered only the data

associated with "complete proteomes" (defined by UniProtKB). The proportion of NPs from all annotated protein sequences is maximal in insects when compared to worms and mammals (Fig. 8.8, right).

8.10.1 A Neuropeptide Precursor Prototype

Studying the molecular processing of NPs is essential for designing a collection of relevant features extracted from sequences. These features should capture the essence of discriminating properties of the "true" *vs.* "false" sets (see Sect. 8.3). The predictor can be used to classify new instances of unknown sequences.

A prototypical example of NPs is the allatostatin family from *A. mellifera* [61]. Approximately 500 neurons in the honeybee brain produce allatostatins [62] which act to inhibit juvenile hormone biosynthesis and reduce food intake. The precursor protein (UniProtKB: P85797, 197 amino acids) produces after cleavage ten active NPs which were identified by MS experiments [61]. The allatostatin NPP of the Pacific beetle cockroach (*Diploptera punctata*) contains 13 identified NPs (Fig. 8.9).

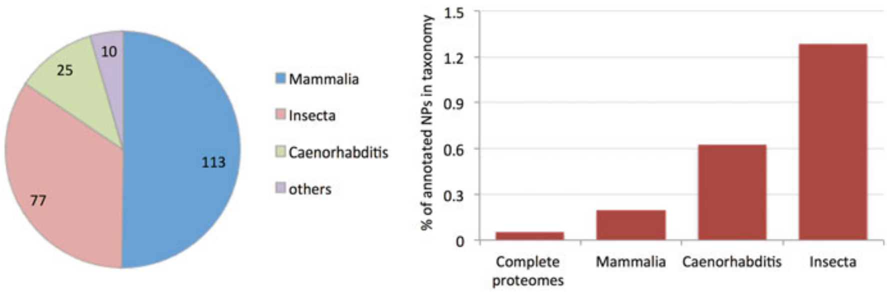


Fig. 8.8 Partition of annotated neuropeptides from major taxonomical groups. Number of annotated neuropeptides (NPs, *left*). Fraction (%) of SwissProt keyword "neuropeptide" from the sequences of "complete proteomes" (*right*). The fraction of NPs from insects is 6.7-fold relative to the fraction of NPs in mammals

```
66- 97    KRL....YDFG..........LG..............KRA..YsyvSEYKRL.........pvYN..FGLGKR
98- 120   SKM....YGFG..........LG..............KR.......DG..RM...........YS..FGLGKR
121-164   DYD....Y.YGeededdqqaIGdedieesdvgdlmdKR..........DRL...........YS..FGLGKR
165-191   ARP....YSFG..........LG..............KRA..P...SGAQRL...........YG..FGLGKR
192-220   GGS...lYSFG..........LG..............KR........GDGRL...........YA..FGLGKRPVNS
221-253   GRSsgsrFNFG..........LG..............KRS..D...DIDFRE...........LEekFAEDKR
254-316    YPqehrFSFG..........LG..............KREveP...SELEAVrne(25)slhYP..FGIRKL
346-367   RRP....FNFG..........LG..............KRI..P........M...........YD..FGIGKR
```

Fig. 8.9 Sequence of neuropeptide precursor P12764 (ALLS_DIPPU, 370 aa) from *Diploptera punctata* (Pacific beetle cockroach). The repeated nature of the sequence is shown. The repeated segments account for 13 bioactive NPs, called allatostatin-1 to allatostatin-13. The NPs are consecutively colored *red* and *blue*. The dibasic residues (cleavage sites) are highlighted in *yellow*

Though each peptide has a unique sequence, all share the Tyr/Phe-Xaa-Phe-Gly-Leu/Ile-NH2 consensus sequence. Furthermore, each active NP is amidated on the terminal Leu/Ile. Evidently, the above properties cannot be captured by methods for remote homology detection that are based on sequence alignments.

8.10.2 Neuropeptide Precursors: Feature Extraction

Similar to the arguments raised for identification of TOLIPs (Sects. 8.6, 8.7, 8.8, and 8.9), Fig. 8.9 illustrates the difficulty in using sequence alignment for identifying NPPs. We thus set out to train a predictor using supervised machine learning. To this end, we compiled nonredundant "positive" and "negative" sets. The nonredundant "positives" included all annotated sequences from SwissProt as well as the automatically inferred sequences from UniProtKB. The "negative" sets included sequences that are basic in nature (i.e., enriched with basic residues) as well as randomly selected proteins from Metazoans with an identical length distribution.

A characteristic "feature" for the majority of NPs is their production from larger precursor proteins (NPP) [63]. In most cases, NPPs produce different NPs that may participate in executing a behavior [64]. A dominant feature is the presence of clusters of dibasic residues that specify these proteolytic cleavage sites. Nevertheless, some NPPs do not use dibasic residues as a cleavage signal.

The goal of extracted features from the training sets is to capture the particular traits and variance between the "positive" and "negative" sets. The information collected to construct a predictor covers:

(A) Biophysical quantitative properties [65] including: (1) the length and molecular weight, (2) frequency of the amino acids or their grouping (e.g., charged amino acids) and dipeptide frequencies (400 features), and (3) quantitative indices, such as aromaticity, instability index, hydropathy, and PI (i.e., isoelectric point).
(B) Binary features that capture the nonrandomized appearance of certain amino acids in short, overlapping windows. This features grouping stems from the occurrence of certain residues near known cleavage sites such as G-KR (Gly, Lys, Arg), lack of flanking proline at cleavage sites, and structural considerations (e.g., disordered, accessible regions).
(C) Appearance and frequency of known sequence motifs. The most important motifs stem from conservation by the processing endopeptidases, such as flanking pairs of basic residues [66]. In addition, we considered potential amidation, hydroxylation, and N-glycosylation sites.
(D) Information-based statistics. The intuition is to trace the entropy, the autocorrelation of the potential cleavage sites, and the repeated nature of the sequences (see an example in Fig. 8.9).

8.10.3 *Prediction of Insect Neuropeptide Precursors*

Testing the performance of the machine learning approach for identifying known and novel NPPs was carried out using a cross-validation (CV) protocol. Accordingly, a substantial fraction of the data (i.e., 10–40 %) was removed and excluded during the training phase and used as a test set. The protocol is repeated multiple times using a different subset of the data each time. The results of the CV tests for each of the NPP candidates were summed up to estimate the accuracy, sensitivity, precision, and AUC (area under ROC curve). The accuracy rate for NPP identification reached a level of 82–89 % for insect NPPs. The class of random forest [17] ensemble decision tree method performed best. Slightly lower performance was recorded for gradient boosting decision trees and linear and nonlinear SVM models [29].

By increasing the thresholds of the prediction tools, we filtered the number of NPP candidates to a few tens. For example, the random forest protocol at a "certainty" threshold of 0.99 reduced the predictions for *B. mori* from ~4000 to only 16. All NPPs are secreted proteins, and thus each has a signal peptide sequence in the N-terminal which is removed prior to the production of the precursor protein. This is a strong feature that was used to remove many of the false positives of the prediction machine.

8.10.4 *Identifying Candidate NPPs in Insect Proteomes*

Two hundred ninety-seven proteins (total of 20,600 protein sequences) in *A. pisum* include a signal peptide (SP) and are thus candidates for being NPPs. A test of our NPP machine learning platform (NeuroPID) yielded about one-third as potential candidates and 13 as high-probability NPPs. Experimental information on these sequences is lacking. The ETH (ecdysis-triggering hormone) precursor was identified among these poorly characterized predicted proteins (Table 8.3). In most insect species, the ETH precursor produces two active peptides [67]. ETH genes and their receptors have also been identified in tick (Arachnida) and water flea (Crustacea). Notably, several of these sequences, while marked as uncharacterized, are highly expressed (Table 8.3). The task of validating these as NPPs calls for functional experimentation and independent evidence (e.g., using MS).

8.11 Insect Short Active Peptides for Human Health and Agriculture

The efficiency and quality of experimentally validated proteins lag behind the explosive growth in sequencing. We expect that the analysis presented in this study will be useful for leveraging the expansion of protein space. In this chapter, we

Table 8.3 Top predictions of neuropeptide precursors from *Acyrthosiphon pisum*

Protein name[a]	Function/expression	Domains[b]
Chemosensory protein-like prec	Pheromone-BP	
Ecdysis-triggering hormone prepro	ETH novel NPP	
Miple protein prec	GF, heparin binding	PTN/MK, C-ter.
Mitochondrial TIM14-like prec	Chaperone -HSP70	DNAJ
Odorant-binding protein 7 prec	GPCR BP	
Odorant-binding protein 8 prec	GPCR BP	
UC protein LOC100159063 prec	Expression—high	
UC protein LOC100161501 prec	Expression—low	
UC protein LOC100162497 prec	Expression—high	TMEMB_9
UC protein LOC100166422 prec	Expression—low	
UC protein LOC100169149 prec	Highly conserved	
UC protein LOC100302326 prec	Expression—medium	
UC protein LOC100574827 prec		

[a]*Prepro* preprotein, *prec* precursor, *UC* uncharacterized, *GF* growth factor
[b]Domains are listed according to the Pfam abbreviations

introduced machine learning approaches for large-scale protein classification of short peptides that resemble animal toxins as well as NPPs and cell modulators.

The therapeutic potential of toxins has been realized and has led to the development of toxin-based drugs, with ICI toxins being the lead for such development [68]. Insect peptides that act in the innate immune system (e.g., defensins) and antibacterial proteins [69] are additional classes of potential drugs which can also be developed as pesticides. Below we outline some of the benefits and applications of these molecules for human health and agriculture.

There are several benefits for the pharmaceutical industry to focus on biological active peptides in general and on toxins and TOLIPs in particular. For example, the 3D high stability of the backbones makes them appealing for drug design, and several toxin-based drugs are already available on the market in synthetic form. A well-known example takes advantage of the mimicry of the MVIIA ω-conotoxin from the marine cone snail *Conus magus* which acts as a blocker of the voltage-gated Ca^{2+} channel [70]. The clinical application of this drug is for chronic, uncontrollable pain. Utilization of the classifiers described earlier in this chapter such as ClanTox and NeuroPID may expand the range of known insect modulators. ProFET (Protein Feature Engineering Toolkit) is another such framework for the machine learning approach to protein function, offering an easy to use, universal platform as well as state of the art results in classification of high-level functions [72].

Another benefit for drug design is that most toxins and the TOLIPs are resistant to proteolysis. This is not only a by-product of their structural compactness but also because many TOLIPs are actually protease inhibitors [56]. This property ensures stability in use as a drug, which is reflected in the protein half-life. Posttranslational modifications on most of these toxins provide an additional layer of stability in the cell and most importantly in the extracellular space.

Insects populate many ecological niches and some of them are considered pest species. Management of pests has significant economic implications. With the increase in environmental awareness, new insecticidal compounds must be explored. The coevolution of insects and plants for millions of years argues that hundreds of TOLIPs are attractive candidates for screening novel targets in plants and animals. Potentially, different folds of TOLIPs can be used in a rational design for as yet unknown targets. Through mimetic approaches, the scaffold of these short proteins can be reduced and directed toward protecting diverse hosts from pests by disturbing and damaging selected membranes of pathogens and even altering mating behavior to control the balance in the ecosystem.

Novel NPPs from insects can play a biotechnological lead in regulating social behavior, metabolic status, and communication. In this view, an exciting genomic initiative with the goal of sequencing 5000 arthropod genomes was recently announced [71]. The expectation is that prediction methods for short proteins will make a valuable contribution for identifying unexplored modes of insect communication, among other features. The power of our method increases with the increase in sequenced genomes, transcriptomes, and proteomes of related species. We expect that the analysis presented in this study will be useful in leveraging future expansions of protein space.

Acknowledgments We thank Noam Kaplan for his seminal contribution to the work. Many of the ideas and formulation of the problem resulted from many stimulating discussions. N.K.'s beautiful thesis was used as a starting point for this chapter. The development and maintenance of ClanTox, NeuroPID, and the protein database have been a long-term effort of many people over the last 10 years. We thank Solange Karsenty for her contribution to the development of ClanTox and NeuroPID platforms and the system group in the School of Computer Science and Engineering for long-term support of our web servers.

References

1. Loewenstein Y et al (2009) Protein function annotation by homology-based inference. Genome Biol 10:207
2. Sasson O, Kaplan N, Linial M (2006) Functional annotation prediction: all for one and one for all. Protein Sci 15:1557–1562
3. Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402
4. Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12:85–94
5. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins 20:216–226
6. Shachar O, Linial M (2004) A robust method to detect structural and functional remote homologues. Proteins 57:531–538
7. Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202
8. Radivojac P et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227

9. Schuldiner S, Shirvan A, Linial M (1995) Vesicular neurotransmitter transporters: from bacteria to humans. Physiol Rev 75:369–392
10. Biegert A, Soding J (2009) Sequence context-specific profiles for homology searching. Proc Natl Acad Sci U S A 106:3770–3775
11. Eddy SR (2009) A new generation of homology search tools based on probabilistic inference. Genome Inform (International Conference on Genome Informatics) 23:205–211
12. Punta M et al (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301
13. Portugaly E, Linial N, Linial M (2007) EVEREST: a collection of evolutionary conserved protein domains. Nucleic Acids Res 35:D241–D246
14. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585
15. Rao HB, Zhu F, Yang GB, Li ZR, Chen YZ (2011) Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. Nucleic Acids Res 39:W385–W390
16. Chapelle O (2007) Training a support vector machine in the primal. Neural Comput 19:1155–1178
17. Breiman L (2001) Random forests. Mach Learn Cybern 45:5–32
18. Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res 31:3692–3697
19. Leslie C, Eskin E, Noble WS (2002) The spectrum kernel: a string kernel for SVM protein classification. Pac Symp Biocomput 2002:564–575
20. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S (2007) Machine learning and its applications to biology. PLoS Comput Biol 3:e116
21. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G (2008) Support vector machines and kernels for computational biology. PLoS Comput Biol 4:e1000173
22. Rappoport N, Linial N, Linial M (2013) ProtoNet: charting the expanding universe of protein sequences. Nat Biotechnol 31:290–292
23. Frith MC et al (2006) The abundance of short proteins in the mammalian proteome. PLoS Genet 2:e52
24. Kondo T et al (2010) Small peptides switch the transcriptional activity of Shavenbaby during Drosophila embryogenesis. Science 329:336–339
25. Ponting CP, Belgard TG (2010) Transcribed dark matter: meaning or myth? Hum Mol Genet 19:R162–R168
26. Lubec G, Afjehi-Sadat L (2007) Limitations and pitfalls in protein identification by mass spectrometry. Chem Rev 107:3568–3584
27. Wu CH (2006) Bioinformatics for proteomics at the Protein Information Resource (PIR). Mol Cell Proteomics 5:S341–S341
28. Rappoport N, Fromer M, Schweiger R, Linial M (2010) PANDORA: analysis of protein and peptide sets through the hierarchical integration of annotations. Nucleic Acids Res 38:W84–W89
29. Ofer D, Linial M (2013) NeuroPID: a predictor for identifying neuropeptide precursors from metazoan proteomes. Bioinformatics. 30(7):931–940.
30. Fry BG (2005) From genome to "venome": molecular origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences and related body proteins. Genome Res 15:403–420
31. Mouhat S, Jouirou B, Mosbah A, De Waard M, Sabatier JM (2004) Diversity of folds in animal toxins acting on ion channels. Biochem J 378:717–726
32. Norton RS, Pallaghy PK (1998) The cystine knot structure of ion channel toxins and related polypeptides. Toxicon 36:1573–1583
33. Terlau H, Olivera BM (2004) Conus venoms: a rich source of novel ion channel-targeted peptides. Physiol Rev 84:41–68
34. Ibanez-Tallon I et al (2002) Novel modulation of neuronal nicotinic acetylcholine receptors by association with the endogenous prototoxin lynx1. Neuron 33:893–903

35. Chimienti F et al (2003) Identification of SLURP-1 as an epidermal neuromodulator explains the clinical phenotype of Mal de Meleda. Hum Mol Genet 12:3017–3024
36. Schoofs L, Beets I (2013) Neuropeptides control life-phase transitions. Proc Natl Acad Sci U S A 110:7973–7974
37. Karsenty S, Rappoport N, Ofer D, Zair A, Linial M (2014) NeuroPID: a classifier of neuropeptide precursors. Nucleic Acids Res 42:W182–W186
38. Brain SD, Cox HM (2006) Neuropeptides and their receptors: innovative science providing novel therapeutic targets. Br J Pharmacol 147(Suppl 1):S202–S211
39. Nassel DR (2002) Neuropeptides in the nervous system of Drosophila and other insects: multiple roles as neuromodulators and neurohormones. Prog Neurobiol 68:1–84
40. Vanden Broeck J (2001) Neuropeptides and their precursors in the fruitfly, Drosophila melanogaster. Peptides 22:241–254
41. Naamati G, Askenazi M, Linial M (2009) ClanTox: a classifier of short animal toxins. Nucleic Acids Res 37:W363–W368
42. Dimmer EC et al (2012) The UniProt-GO Annotation database in 2011. Nucleic Acids Res 40:D565–D570
43. Kim BY et al (2013) Antimicrobial activity of a honeybee (*Apis cerana*) venom Kazal-type serine protease inhibitor. Toxicon 76:110–117
44. Palmiter RD (1998) The elusive function of metallothioneins. Proc Natl Acad Sci U S A 95:8428–8430
45. Tian C et al (2008) Gene expression, antiparasitic activity, and functional evolution of the drosomycin family. Mol Immunol 45:3909–3916
46. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786
47. Lipkind GM, Fozzard HA (1994) A structural model of the tetrodotoxin and saxitoxin binding site of the Na+channel. Biophys J 66:1–13
48. Kaplan N, Morpurgo N, Linial M (2007) Novel families of toxin-like peptides in insects and mammals: a computational approach. J Mol Biol 369:553–566
49. Kloog Y et al (1988) Sarafotoxin, a novel vasoconstrictor peptide: phosphoinositide hydrolysis in rat heart and brain. Science 242:268–270
50. Sousa SR, Vetter I, Lewis RJ (2013) Venom peptides as a rich source of cav2.2 channel blockers. Toxins 5:286–314
51. Su M, Ling Y, Yu J, Wu J, Xiao J (2013) Small proteins: untapped area of potential biological importance. Front Genet 4:286
52. Nijhout HF, Grunert LW (2010) The cellular and physiological mechanism of wing-body scaling in *Manduca sexta*. Science 330:1693–1695
53. Mizoguchi A et al (2013) Prothoracicotropic hormone acts as a neuroendocrine switch between pupal diapause and adult development. PLoS One 8:e60824
54. Schofield CJ, Jannin J, Salvatella R (2006) The future of Chagas disease control. Trends Parasitol 22:583–588
55. Lee PY, Wang JX, Parisini E, Dascher CC, Nigrovic PA (2013) Ly6 family proteins in neutrophil biology. J Leukoc Biol 94:585–594
56. Tirosh Y, Ofer D, Eliyahu T, Linial M (2013) Short toxin-like proteins attack the defense line of innate immunity. Toxins 5:1314–1331
57. Naamati G, Askenazi M, Linial M (2010) A predictor for toxin-like proteins exposes cell modulator candidates within viral genomes. Bioinformatics 26:i482–i488
58. Cui L, Webb BA (1996) Isolation and characterization of a member of the cysteine-rich gene family from Campoletis sonorensis polydnavirus. J Gen Virol 77(Pt 4):797–809
59. Jekely G (2013) Global view of the evolution and diversity of metazoan neuropeptide signaling. Proc Natl Acad Sci U S A 110:8702–8707
60. Insel TR, Young LJ (2000) Neuropeptides and the evolution of social behavior. Curr Opin Neurobiol 10:784–789
61. Hummon AB et al (2006) From the genome to the proteome: uncovering peptides in the Apis brain. Science 314:647–649

62. Kreissl S, Strasser C, Galizia CG (2010) Allatostatin immunoreactivity in the honeybee brain. J Comp Neurol 518:1391–1417
63. Mirabeau O et al (2007) Identification of novel peptide hormones in the human proteome by hidden Markov model screening. Genome Res 17:320–327
64. Mentlein R, Dahms P (1994) Endopeptidases 24.16 and 24.15 are responsible for the degradation of somatostatin, neurotensin, and other neuropeptides by cultivated rat cortical astrocytes. J Neurochem 62:27–36
65. Artimo P et al (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40:W597–W603
66. Southey BR, Sweedler JV, Rodriguez-Zas SL (2008) Prediction of neuropeptide cleavage sites in insects. Bioinformatics 24:815–825
67. Roller L et al (2010) Ecdysis triggering hormone signaling in arthropods. Peptides 31:429–441
68. Fox JW, Serrano SM (2007) Approaching the golden age of natural product pharmaceuticals from venom libraries: an overview of toxins and toxin-derivatives currently involved in therapeutic or diagnostic applications. Curr Pharm Des 13:2927–2934
69. Lai Y, Gallo RL (2009) AMPed up immunity: how antimicrobial peptides have multiple roles in immune defense. Trends Immunol 30:131–141
70. Brady RM, Baell JB, Norton RS (2013) Strategies for the development of conotoxins as new therapeutic leads. Mar Drugs 11:2293–2313
71. Consortium iK (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. J Hered 104:595–600
72. Ofer, Dan, and Michal Linial. "ProFET: Feature engineering captures high-level protein functions." Bioinformatics (2015): btv345.